MATHEMATICAL AND COMPUTATIONAL EVOLUTIONARY BIOLOGY

May 26-30, 2019
Porquerolles, France

LIRMM

Institut Pasteur

C3BI USR 3756

cnrs

LabEx NUMEV

INRA SCIENCE & IMPACT

CeMEB

MUSE MONTPELLIER UNIVERSITY of EXCELLENCE

EcoStat GdR

GdR BiM

INFORMATION

Meeting Point

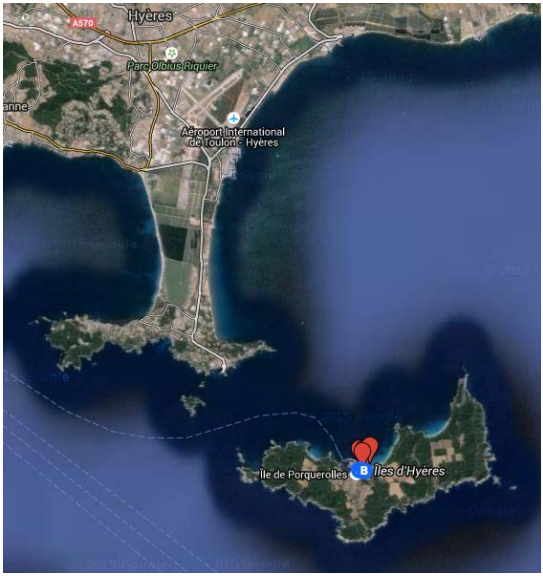
Directly at IGESA Center for the welcoming drink at « le Théâtre de verdure » (next to reception) on Sunday, May 26th.

However, you can arrive at IGESA and get the key from 4pm
Aim to get the 6.30pm ferry (or earlier) at La Tour Fondue.

In case of problems :
Olivier Gascuel : +33 (0) 06 48 12 14 82
Fabio Pardi : +33 (0)6 83 23 20 14
Krister Swenson : +33 (0)7 81 71 60 90
Stephane Guindon : +33(0)7 83 61 69 06.
=> inside France use 0 to start (06 48...), outside france use +33 without the 0 (336 48...)

Location

The conference will be held at the IGESA centre, in the village of Porquerolles, five minutes from the harbour and the beach. It is located in a nature reserve, the Port-Cros National Park, one of the most intact coastal areas in the Mediterranean.



Practical INFORMATION

Hôtel Club IGESA

Rue de la Douane
Ile de Porquerolles
83 400 HYERES
Tel : 04 94 12 31 80

Coordonnées GPS :
43.000402,6.205076
WI-FI : available
but very limited speed

Ferry :

TLV-TVM : 04 94 58 21 81
Return ticket: 19,50€
La Tour Fondue -> Porquerolles :
7:30, 9:00, 9:30, 10:00, 10:30, 11:00, 11:30, 12:00, 12:30, 13:30, 14:30, 15:30, 16:30, 17:30, 18:30
Porquerolles -> La Tour Fondue :
7:00, 8:30, 9:30, 10:00, 10:30, 11:00, 11:30, 14:00, 15:00, 16:00, 17:00, 18:00, 19:00
After hours shuttle :
One way ticket : 18,50€ (buy your ticket directly on board)
La Tour Fondue -> Porquerolles : 19 :45, 23 :15
Porquerolles -> La Tour Fondue : 20:00, 23:30

Taxi-Boat:

Taxi Boat « Le Pelican » : 04 94 58 31 19
The taxi-boat costs 16.50€ per person, when at least 6 people are booked on it.

Car Park on Tour Fondue:

Cars are not allowed on the island, so if you come by car you'll have to leave it in a car park at La Tour Fondue. You can book a place.
Car Park Porquerolles (13h/24h :15€)
Car park des Iles (videosurveillance) : 04 94 58 90 78 (24h : 15€)
Car park Indigo (watchman 24/24) : 04 94 01 99 28 (24h : 17,20€)

Bus Hyères-La Tour Fondue:

If you come by train or plane to Hyeres, you can take the bus line 67 on «Reseau Mistral»
More information is available with your mobile : http://m.reseaumistral.com/
Or call a taxi : Taxi Hyères 04 94 00 60 00

Program

Sunday, May 26th

19:00 : Welcoming drink - Théatre de Verdure IGESA

20:00 : Dinner

Monday, May 27th

09h00 : Welcome
09h15 – 10h30 : **Dannie Durand** (Carnegie Mellon University, USA)
«Genome Evolution»

10h30 : Coffee break

11h00 – 12h00 : 3X20mn min TALKS (including questions)
Cédric NotreDame : ***Fast and accurate large multiple sequence alignments using root-to-leave regressive computation***
Jean Cury : ***Back to the future of bacterial population genomics***
Kimberly Gilbert : ***Disentangling the complex interactions of background selection, associative overdominance, and recombination in determining genomic diversity***

12h20 : Lunch

14h00 – 15h00 : 3x20 min TALKS (including questions)
Nicola De Maio : ***Fast and accurate statistical evolutionary alignment***
Elise Kerdoncuff : ***A fast genome chopper to detect strong species decline.***
Jakub Voznica : ***Deep Learning for Parameter Inference in Phylodynamics***

15h00 : Break

15h15 – 16h30 : KEYNOTE
Andrew Francis (Centre for Research in Mathematics, Western Sydney University, AU)
The Mathematics of Phylogenetics Networks

16h30 - 18h30 : Freetime, beach...

18h30-20h00 : POSTERS
Wine and discussion (posters 1-10)

20h00 : Dinner

Program

Tuesday, May 28th

09h15 – 10h30 : KEYNOTE
Michael Blum (TIMC, CNRS and University Grenoble Alpes, FR)
Genome wide association studies and polygenic models

10h30 : Coffee break

11h00 – 12h20 : 4X20 min TALKS (including questions)
Sophia Lambert : ***Estimating diversification rates from phylogenies when the total number of taxa is unk***nown
Lars Jermiin : ***Detection of tree-likeness in phylogenetic data***
Sebastian Duchene : ***Infectious disease phylodynamics using genomic and notification data***
Harald Ringbauer : *Inferring runs of homozygosity from low coverage ancient DNA data*

12h20 : Lunch

14h00 - 20h00: Free afternoon (beach, hiking, theorems, etc)

20h00 : Dinner

Wednesday, May 29th

09h15 – 10h30 : KEYNOTE
Laura Eme (ETTEMA LAB Microbial diversity and evolution, Uppsala University, SE)
Phylogenomics for the origin and early evolution of eukaryotes

10h30 : Coffee break

11h00 – 12h00 : 3x 20 min TALKS (including questions)
Marc Manceau : ***Inferring the ancestral population size under birth-death models***
Miguel Navascues : ***Joint Inference of Demography and Selection from Genomic Temporal Data Using Approximate Bayesian Computation***
Veronika Boskova : ***Phylodynamic inference from large datasets with many duplicate sequences***

12h00 : Lunch

Program

14h00 – 15h00 : 3x20 min TALKS (including questions)
Carolin Kosiol *: IQ-TREE-POMO: Polymorphism-aware tree estimation*
Yan Wong : *Inferring the ancestry of everyone*
Jonathan Mitchell : *Testing n-Taxon Species Trees with the Multispecies Coalescent Model*

15h00 - 15h15 : Break

15h15 – 16h30 : KEYNOTE
Olivier Delaneau (University of Lausanne, FR)
Haplotype estimation with sub-linear complexity

16h30 - 18h30 : Freetime, beach...

18h30 - 20h00 : POSTERS
Wine and discussion (posters 11-22)

20h00 : Dinner

Thursday May 30th

09h30 – 10h30 : 3x20mn TALKS (including questions)
Donate Weghorn : *Probabilistic approaches to positive and negative selection inference on coding regions in cancer*
Maryam Alamil : *A statistical learning approach to infer transmissions of infectious diseases from deep sequencing data*
Cédric Chauve : *MLST genotyping of bacterial pathogens using whole-genome sequencing data*

10h30 : Coffee break

11h00 – 12h15 : KEYNOTE
Guy Sella (Department of Biological Sciences, Center for Computational Biology and Bioinformatics, Program for Mathematical Genomics, Columbia University, New York, USA)
Advances in Population Genetics

12h15 : Lunch and farewell session

14h00 : First ferry to « La Tour Fondue »

Keynote speakers

Keynote speakers

> Michael BLUM
TIMC, CNRS and University Grenoble Alpes, FR
Genome wide association studies and polygenic models

> Olivier DELANEAU
University of Lausanne, FR
Haplotype estimation with sub-linear complexity

> Dannie DURAND
Carnegie Mellon University, USA
Genome Evolution

> Laura EME
ETTEMA LAB Microbial diversity and evolution,
Uppsala University, SE
Phylogenomics for the origin and early evolution of eukaryotes

> Andrew FRANCIS
Centre for Research in Mathematics, Western Sydney University, AU
The Mathematics of Phylogenetics Networks

> Guy SELLA
Department of Biological Sciences, Center for Computational Biology and Bioinformatics, Program for Mathematical Genomics, Columbia University, New York, USA
Advances in Population Genetics

Edgar Garriga[1,2];Paolo Di Tommaso[1]; Cedrik Magis[1,2]; Ionas Erb[1]; Hafid Laayouni[3,4]; Fyodor Kondrashov[5]; Evan Floden[1]; **Cedric Notredame**[1,2]
[1] Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain. [2] Universitat Pompeu Fabra (UPF), Dr. Aiguader 88, Barcelona 08003, Spain. [3] Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Barcelona, Catalonia, 10 Spain. [4]Bioinformatics Studies, ESCI-UPF, Barcelona, Spain. [5]Institute of Science and Technology, Klosterneuburg, Austria.

Fast and accurate large multiple sequence alignments using root-to-leave regressive computation

Inferences derived from large multiple alignments of biological sequences are critical to many areas of biology, including evolution, genomics, biochemistry, and structural biology. However, the complexity of the alignment problem imposes the use of approximate solutions. The most common is the progressive algorithm, which starts by aligning the most similar sequences, incorporiating the remaining ones following the order imposed by a guide-tree. We developed and validated on protein sequences a regressive algorithm that works the other way around, aligning first the most dissimilar sequences. Our algorithm produces more accurate alignments than non-regressive methods, especially on datasets larger than 10,000 sequences. By design, it can run any existing alignment method in linear time thus allowing the scale-up required for extremely large genomic analyses.

Jean Cury[1], Ben Haller[2], Théophile Sanchez[1], Guillaume Charpiat[1], Flora Jay[1]
[1] Laboratoire de Recherche en Informatique (LRI), Orsay, France [2] Department of Biological Statistics and Computational Biology, Cornell University, USA

Back to the future of bacterial population genomics

Population genomics inferences are usually performed using simulation-based approaches such as Approximate Bayesian Computation (ABC) methods to test for demographic patterns under a given model. Alternatively, other approaches are simulation-free, such as skyline plots, which aim at inferring the coalescent rates from sequence data and using them for demographic inference. In bacterial population genomics, the latter method has been widely applied to analysis of pathogenic bacteria. However, it has been shown that violation of standard neutral assumptions can lead to incorrect inferences from skyline plots (Lapierre et al., 2016). Despite such problems with analytical methods, simulation-based approaches are not often used in bacterial population genomics, probably due to the lack of powerful bacterial population genomics simulators. To our knowledge, there is no bacterial population genomics simulator that: (1) is computationally efficient; (2) can handle both demography and selection; (3) implements bacterial recombination that occurs through horizontal gene transfers of homologous DNA material. We will present how we implemented a bacterial simulator using SLiM, a forward-time simulator with a scripting interface (Haller et al., 2018). Data can be generated under the Wright-Fisher model, or with a non-Wright–Fisher model that gives greater control at the individual level. We will also highlight various features of SLiM that can be used to create complex bacterial population genomics scena-

rios. We used our simulator to generate the largest dataset of bacterial haplotypes so far created. We will present preliminary work on selection inference under various demographical models without using summary statistics, using a deep learning framework. Deep learning methods are computationally costly to train, but this step is done one time on specific hardware and the model can then be used for prediction on any machine at low computational cost. We designed specific neural network architectures to take into account the structure of the data during training, and we are using cutting-edge techniques to investigate selection signal along the chromosome. Finally, we will discuss drawbacks and challenges in deep learning when applied to population genomics. M. Lapierre, C. Blin, A. Lambert, G. Achaz, and E. P. C. Rocha, "The Impact of Selection, Gene Conversion, and Biased Sampling on the Assessment of Microbial Demography," Mol. Biol. Evol., 2016.B. C. Haller and P. W. Messer, "SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model," Mol. Biol. Evol., 2019.

Kimberly J. Gilbert[1,2]; Fanny Pouyet[1,2]; Stephan Peischl[2,3]; Laurent Excoffier[1,2]
[1] Institute of Ecology and Evolution, University of Bern, Bern, Switzerland; [2] Swiss Institute of Bioinformatics, Lausanne, Switzerland; [3] Interfaculty Bioinformatics Unit, University of Bern, Bern, Switzerland

Disentangling the complex interactions of background selection, associative overdominance, and recombination in determining genomic diversity

Understanding the factors that contribute to the generation and maintenance of genetic diversity is a core goal of evolutionary biology. In order to make proper demographic inference or properly identify loci under selection, we must understand the processes contributing to neutral genetic diversity. Natural selection acting on deleterious variants has large potential to influence changes in diversity of neutral variants. Two key non-selective processes, background selection (BGS) and associative overdominance (AOD), impact neutral genetic diversity in complex and potentially opposing ways. It is well established that BGS reduces genetic diversity through the linkage of neutral alleles with deleterious variants subject to purifying selection. Conversely, AOD can preserve genetic diversity when neutral variants are linked to recessive deleterious variants. Recessive deleterious variants are only subject to selection as homozygotes and can therefore remain at low frequency as heterozygotes in the population without being eliminated by negative selection. When these loci are linked to neutral variants, many rare or singleton neutral variants can be preserved in the population, particularly in regions of low recombination. Previous work has investigated the interactions of BGS and AOD in driving neutral genetic diversity (Charlesworth et al. 2009, Nordborg & Charlesworth 1996) using either single-locus models or models lacking recombination. Here we expand these investigations to include a range of selection coefficients and dominance parameters to test the impact of BGS and AOD on genetic diversity under multi-locus scenarios with varying recombination rates across the genome. Interestingly, our results highlight the large impact of selective interference from linkage between deleterious variants, effectively changing the strength of selection against linked loci in the genome. Observed genetic diversity depends highly on the combination of all parameters (s, h, and r), with different processes driving the various results. Our results highlight the fact that the maintenance and/or generation of neutral genetic diversity depends highly on the genetic architecture and distribution of fitness effects

of mutations in the genome. Furthermore, this emphasizes the importance of identifying neutral loci whose evolution is not constrained by selective forces in order to make accurate evolutionary inferences (Pouyet et al 2018). Charlesworth B, Betancourt AJ, Kaiser VB, Gordo I (2009) Genetic recombination and molecular evolution. Cold Spring Harbor Symposia on Quantitative Biology, vol. LXXIV.Nordborg M, Charlesworth B (1996) The effect of recombination on background selection. Genet. Res. 67: 159-174.Pouyet F, Aeschbacher S, Thiéry A, Excoffier L (2018) Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. eLife, 7:e36317.

Nicola De Maio[1]
[1] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, CB10 1SD, United Kingdom

Fast and accurate statistical evolutionary alignment

Sequence alignment is essential for phylogenetic and molecular evolution inference, as well as in many other areas of bioinformatics and evolutionary biology. Inaccurate alignments can lead to severe biases in most downstream statistical analyses. Statistical alignment based on probabilistic models of sequence evolution addresses these issues by replacing heuristic score functions with evolutionary model-based probabilities. However, score-based aligners and fixed-alignment phylogenetic approaches are still more prevalent than methods based on evolutionary indel models, mostly due to computational convenience. Here, I present new techniques for improving the accuracy and speed of statistical evolutionary alignment. The "cumulative indel model" approximates realistic evolutionary indel dynamics using differential equations. "Adaptive banding" reduces the computational demand of most alignment algorithms without requiring prior knowledge of divergence levels or pseudo-optimal alignments. Using simulations, I show that these methods lead to faster and more accurate parameter and pairwise alignment inference. The cumulative indel model and adaptive banding can therefore improve the performance of alignment and phylogenetic methods.

Jakub Voznica[1,2]; Anna Zhukova[1]; Tristan Dot[1]; Kary Ocaña[1]; Frédéric Lemoine[1]; Mathieu Moslonka-Lefebvre[1]; Olivier Gascuel[1]
[1] Unité Bioinformatique Evolutive, Institut Pasteur, C3BI – USR 3756 IP & CNRS, Paris, France; [2] Université Paris Descartes, Sorbonne Paris Cité, Paris, France

Deep Learning for Parameter Inference in Phylodynamics

Phylodynamics is a meeting point between phylogeny and epidemiology. In this rising field, phylogenetic trees are combined with mathematical epidemiology models to estimate epidemiological parameters. Phylogenetic trees are reconstructed using viral genetic data sampled from infected patients. By studying such trees, we gain insight into viral evolution and transmission dynamics. Several standard models from epidemiology (e.g. susceptible-infectious-recovered or SIR) were translated into phylodynamic ones, and enabled to extract information by estimating epidemiological parameters (e.g. basic reproductive number R0) from phylogenies inferred using molecular data. Similar models and approaches were also used successfully in ecology, to study macro-evolution and speciation processes.Standard phylodynamic methods include maximum

likelihood and bayesian approaches, which both need to be set up each time the model changes: analytic formulae to compute the likelihood of the data (tree) are specific to the studied model, do not exist for complex ones, and generally involve highly time-consuming calculations. As an alternative, Approximate Bayesian Computation (ABC) does not require likelihood formulae, as it is based on learning from predefined patterns in the simulated data, so called summary statistics. However, one needs to make sure that the chosen summary statistics convey enough signal for statistical learning methods to estimate model parameters accurately. Thus, if we change the model, we generally need to design new appropriate summary statistics.We tried to bypass the limitations of both techniques by combining concise bijective tree representations using real-valued vectors, with deep learning using neural networks. We propose such a tree representation, with an encoding size that is linear in the tree size. Our experiments with deep neural networks show that the approach is able of automatic feature extraction, and thus does not need to learn from predefined summary statistics. With standard models (e.g. BDM, BDSIR, BDEI, BDsa) where likelihood formulae and efficient implementations are available, the estimation accuracy of the deep learning and likelihood approaches are comparable.This approach thus opens the door to the usage of more complex and innovative phylodynamic models. Furthermore, even if training the neural network is time-consuming, inference itself is extremely fast and the usage of our method could become privileged in the context of real-time surveillance.

Elise Kerdoncuff[1],[2]; Amaury Lambert[2],[3]; Guillaume Achaz[1],[2]
[1] Institut de Systématique, Evolution, Biodiversité (ISYEB), MNHN, Paris, France; [2] Centre Interdisciplinaire de Recherche en Biologie (CIRB), Collège de France, Paris, France; [3] Laboratoire de Probabilités et Modèles Aléatoires (LPMA), UPMC, Paris, France.

A fast genome chopper to detect strong species decline.

Only 5% of described species have a conservation status. Methods used to assess conservation status cannot be generalized to all species. We developed a new method to study demography based on the length of compatible blocks along the genome, a.e. blocks of nucleotides within which recombination events are not detectable. From whole-genome data of multiple individuals in a population, we can chop a chromosome into compatible blocks in seconds. Lengths of compatible blocks depend on the frequency of recombination events which is influenced by the ancestral history of the population. Using the distribution of block lengths, we can discriminate a constant population and a declining one. This method can infer a very recent decline of a population from DNA sequences. It could be a new tool to assess conservation status in a wide range of species.

Sophia Lambert[1,2]; Amaury Lambert[3]; Helene Morlon[1]
[1] Institut de biologie de l'ENS Paris (IBENS), Paris, France; [2] Muséum National d'Histoire Naturelle (MNHN), Paris, France; [3] Centre interdisciplinaire de recherche en biologie (CIRB), Collège de France, Paris, France

Estimating diversification rates from phylogenies when the total number of taxa is unknown

Estimating diversification rates (i.e. rates of speciation and extinction) is crucial to our understanding of life history on Earth. Over the past decades, phylogenetic models of diversification have increasingly been used to obtain such estimates [1]. While current approaches can accommodate missing species in the phylogeny of the studied clade [2], they can only perform when the sampling fraction (i.e. the number of taxa studied given the total number of species) is known. However, for many groups of organisms, such as microorganisms, we are lacking this information. Pioneering studies have analysed diversification rates in microorganisms [3] by using statistical methods for estimating global scale diversity [4], but these have provided rough diversity estimates. As a result, diversification rates are imprecisely estimated for the majority of life on Earth. Here, we develop an approach for estimating diversification rates from phylogenies when global diversity is unknown. We consider a probabilistic model with multiple clades evolving independently according to a birth-death process. In this model, species from each clades are sampled at present independently, but from a common sampling distribution. We then estimate jointly the specific-clade diversification rates and the parameter of the sampling distribution using a Bayesian framework. We test this inference approach using Monte Carlo simulations and show its relevance when trying to estimate the global scale diversity of the clades. Finally, we illustrate our approach using a large diatoms phylogeny - a group of organisms where species diversity remains mostly unknown.We find that our inference approach is able to recover simulated parameters - rates of speciation, extinction and the parameter of the sampling distribution when global diversity is unknown. Furthermore, we characterize the level of diversity estimate uncertainty in which our inference should be used instead of traditional ones. Overall, this study highlights the need to develop dedicated methods to understand the evolutionary dynamics and diversification processes of poorly described clades.
Keywords: diversification rates, unknown species diversity, phylogenetics, Bayesian approach.
[1] Morlon, Hélène. «Phylogenetic approaches for studying diversification.» Ecology letters 17.4 (2014): 508-525.[2] Nee, Sean, Robert Mccredie May, and Paul H. Harvey. «The reconstructed evolutionary process.» Phil. Trans. R. Soc. Lond. B 344.1309 (1994): 305-311.[3] Lewitus, Eric, et al. «Clade-specific diversification dynamics of marine diatoms since the Jurassic.» Nature Ecology & Evolution 2.11 (2018): 1715.[4] Quince, Christopher, Thomas P. Curtis, and William T. Sloan. «The rational exploration of microbial diversity.» The ISME journal 2.10 (2008): 997.

Lars S Jermiin[1][2]
[1]Research School of Biology, Australian National University, Canberra, Australia; [2] School of Biology and Environmental Science, University College Dublin, Belfield, Ireland

Detection of tree-likeness in phylogenetic data

Most model-based phylogenetic methods assume: (a) that the phylogenetic data evolved on a single bifurcating tree, (b) that the evolutionary processes operating at variable sites in the data can be modelled accurately using independent and identically distributed (IID) Markov processes, and (c) that these Markov processes were stationary, reversible and homogeneous (SRH). These general assumptions apply equally to single-gene, concatenated-gene and gene-tree/species-tree approaches in phylogenetics. The tree-likeness assumption is violated when some of the sequences have undergone recombination. However, violation of the assumption is rarely assessed in sufficient detail before phylogenetic analyses are commenced. Failure to do so could have undesirable consequences. In this talk, I propose a simple method to determine whether recombination is

likely to have occurred. It accommodates evolutionary processes that may interfere with the detection of recombination events, so it is unlikely to be affected by factors interfering with the detection of such events. The method is accurate and fast, partly because it is not concerned with the identification of the actual breakpoints in the DNA and partly because it is a non-parametric approach.

Sebastian Duchene[1]
[1] Dept of Microbiology and Immunology Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, 3000, Australia

Infectious disease phylodynamics using genomic and notification data

Genome surveillance is increasingly common for infectious pathogens, ranging from those that cause seasonal outbreaks (e.g. influenza viruses) to those that cause foodborne diseases (e.g. bacteria of the genus Salmonella). Phylodynamic models can take advantage of such data to infer epidemiological dynamics, for example those based on the exponential growth coalescent and the birth-death process. Here we investigate the potential of including notification data in such phylodynamic analyses; that is, information about confirmed cases that have not been sequenced. Using a simulation study, we demonstrate that birth-death models can capitalise on notification data to eliminate bias in estimates of R0 and other epidemiological parameters, particularly when the sampling rate has not been constant over time. Coalescent models are more robust than birth-death models to violations of the sampling scheme, but they are not well suited for notification data, with biases estimates of some parameters. Finally, we analyse genomic and notification data of Human Respiratory Syncytial Virus to give guidelines for including such information in empirical studies.

Harald Ringbauer [1]; John Novembre [1,2]; Matthias Steinrücken [2]
[1] Department of Human Genetics, University of Chicago, Chicago, IL. [2] Department of Ecology and Evolution, University of Chicago, Chicago, IL.

Inferring runs of homozygosity from low coverage ancient DNA data

The ancient DNA revolution has delivered spectacular new insight into population history of humans, and starts to do so also for other organisms. Here I present work on a novel computational method to detect long runs of homozygosity (ROH) for such data. These blocks are the direct genetic signposts of inbreeding. As such, the frequency and length of ROH blocks yields ample insight into recent population history. It is possible to identify ROH in high coverage present-day datasets, by scanning for regions that lack heterozygote markers. But this strategy frequently fails for ancient individuals: The often very low depth (<1x) makes reliable diploid genotype calls impossible for most sites. Our refined method makes use of linkage disequilibrium information from a panel of reference haplotypes under a Hidden Markov Model (HMM). It scans for long stretches of genome that are imperfect copies from single haplotypes in the reference panel. To showcase an application, we apply the method termed HAPSBURG (Haplotype Block Sharing by uninterrupted recent Genealogy) to simulated data and data from ancient humans.

Marc Manceau[1], Timothy Vaughan[1], Tanja Stadler[1]
[1] ETH Zürich

Inferring the ancestral population size under birth-death models

Birth-death models are ubiquitous in evolutionary biology. They are used as an underlying tree prior, or as an underlying population dynamics prior in studies spanning fields as diverse as macroevolution, linguistics, or epidemiology. Computing reliable estimates of the ancestral number of species, languages or infected individuals, i.e. estimates of past population size, is key to the understanding of past processes in these fields.In both macroevolution and epidemiology, these inferences have initially relied on the fossil record and the case counts record, modeled as a sampling of individuals from the full process through time [1]. In the recent years, a huge effort of molecular sequencing, either of present-day species or of pathogens, lead to the reconstruction of phylogenies, which can also be used to get hints on the ancestral population size [2,3].Here, we focus on inferring the past population sizes through time, when we jointly observe a record of sampling times of ancestors, together with the phylogenetic tree of a subsample of the process.Two main approaches have been introduced to tackle this question. The first one consists in computing analytical estimates based on very crude simplifications of the data, such as, e.g, taking only into account the number of individuals at the beginning or at the end of the process [4].The second one consists in performing intensive Monte-Carlo simulations to produce population size trajectories conditioned on the observed data [5]. These Monte-Carlo methods thus produce the most accurate estimates, at the cost of an increased computational burden, thus preventing their use on very big datasets.Here, we present a third approach which allows to compute the law of the population size conditioned on the observed data, in a more efficient way than previously proposed Monte-Carlo algorithms. While not performing a new task, the relative efficiency of our method paves the way towards considering much bigger datasets, or to the extension of the method to multi-type or density-dependent birth-death processes.[1]Starrfelt and Liow, 2016, PTRSB,How many dinosaur species were there? Fossil bias and true richness estimated using a Poisson sampling model.[2]Stadler et al., 2013, PNAS,Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV).[3]Ratmann et al., 2016, MBE,Phylogenetic tools for generalized HIV-1 epidemics: findings from the PANGEA-HIV methods comparison.[4] Morlon et al., 2011, PNAS,Reconciling molecular phylogenies with the fossil record.[5]Vaughan et al., 2018, bioRxiv,Estimating epidemic incidence and prevalence from genomic data.

Vitor A.C. Pavinato**[1,2];** Stéphane de Mita[3]; Jean-Michel Marin[2,4]; **Miguel Navascués**[1,4]
[1] INRA UMR CBGP, Montpellier, France; [2] Université de Montpellier UMR IMAG, Montpellier, France; [3] INRA UMR IAM, Nancy, France; [4] Institut de Biologie Computationnelle (IBC), Montpellier, France

Joint Inference of Demography and Selection from Genomic Temporal Data Using Approximate Bayesian Computation

Most population genetic studies use genotypic or allelic frequency data obtained from several populations sampled at the same time point. However, temporal population genetics data offers a more powerful way of studying complex dynamics, since we can follow allele frequency changes over time. Disentangling the effects of selection and demography is a long-standing difficulty in

population genetics. Recent theoretical works based on simulations have shown that the interaction between the signal of selection bias the demographic inference when selection pervasive. One potential solution is the co-estimation of neutral and selective parameters using simulation-based methods as Approximation Bayesian Computation (ABC). However, traditional ABC approaches are computationally demanding, and their implementation with explicit selection models was unrealistic. The introduction of random forests in ABC reduced the computational burden, making it possible to study complex dynamics with few simulations. We propose the use of ABC Random-Forests to implement the joint inference and co-estimate neutral and selective parameters in temporal population genomics datasets. Our results show that the proposed framework can jointly infer demography and selection, allowing to distinguish true demography (census size) from genetic drift (effective population size), as well as estimate the population genetic load (selection parameter), proportion of loci under selection and classify loci as neutral or adaptive.

Veronika Bošková[1,2,3], Ankit Gupta[2], Tanja Stadler[2,3]
[1] Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University Vienna, Vienna, Austria; [2] Department of Biosystems Science and Engineering, Eidgenössische Technische Hochschule (ETH) Zürich, Basel, Switzerland; [3] SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Phylodynamic inference from large datasets with many duplicate sequences

The larger the dataset, the more detailed information about the population under study it can offer. Deep sequencing approaches allow for precise quantification of sequences in the population. However, the amount of data resulting from the sequencing efforts represents a computational overload for existing Bayesian phylogenetic and phylodynamic methods.Especially in fast evolving and reproducing populations such as RNA viruses, the population of sequences is often very diverse but also very repetitive. Using heuristic approaches of selecting only unique sequences or randomly subsampling the dataset reduces the computational time. Nevertheless, the parameter estimates are less exact, and/or less precise than those resulting from the analysis of the full dataset.We have developed a method for efficient reconstruction of viral dynamics from an alignment of unique sequences and their respective frequencies. By using all the available information, the method is very precise. The method avoids sampling exact topology of duplicate sequences, thereby reducing the computational burden and allowing analysis of around 20'000 sequences within 14 days. The method is implemented as a package for BEAST 2 software [Bouckaert, 2014].We propose a further improvement of the above method by using analytic, rather than numeric, integration over the subtrees of duplicate sequences. This should make the method easily applicable to datasets of several tens of thousands of sequences. References:Bouckaert, Remco, et al. «BEAST 2: a software platform for Bayesian evolutionary analysis.» PLoS Comput Biol 10.4 (2014): e1003537.

Dominik Schrempf [1], Rui Borges[2], Bui Quang Minh [3], and **Carolin Kosiol [2,4]**
[1]Department of Biological Physics, Eötvös Loránd University, Budapest, Hungary; [2] Institut für Populationsgenetik, Vetmeduni Vienna, Wien, Austria; [3] Ecology and Evolution, Research School of Biology, Australian National University, Canberra, Australia; [4] Centre for Biological Diversity, University of St Andrews, St Andrews, United Kingdom

IQ-TREE-POMO: Polymorphism-aware tree estimation

Molecular phylogenetics has neglected polymorphisms within present and ancestral populations for a long time. Recently, multispecies coalescent based methods have increased in popularity, however, their application is limited to a small number of species and individuals. We have introduced a polymorphism-aware phylogenetic model (PoMo), which overcomes this limitation and scales well with the increasing amount of sequence data. PoMo circumvents handling of gene trees and directly infers species trees from allele frequency data.PoMo extends any DNA substitution model and additionally accounts for polymorphisms in the present and in the ancestral population by expanding the state space to include polymorphic states. It is a selection-mutation model which separates the mutation process from the fixation process. PoMo naturally accounts for incomplete lineage sorting because ancestral populations can be in a polymorphic state. Our method can accurately and time-efficiently estimate the parameters describing evolutionary patterns for phylogenetic trees of any shape (species trees, population trees, or any combination of those). We have implemented our PoMo approach as software package IQ-TREE-POMO with several new features: (i) a search for the statistically best-fit mutation model (ModelFinder), (ii) the ability to allow mutation rate variation across sites (e.g., gamma distribution), assessment of branch support values (bootstrapping and jackknifing), (iv) simulator of sequences evolving under PoMo (bmm-simulate), and (v) inference of allelic selection.Applications using great ape data sets will be presented. In particular, the new genome-wide data set of seven baboon populations (genus Papio) present a unique opportunity to apply our method to a primate clade that involves more complex processes than those usually assumed by phylogenetic models. The history of Papio includes episodes of introgression or admixture among genetically distinct lineages. We will discuss the effect of this complex history on genome-wide phylogenetic inference with PoMo as well as other approaches.

Yan Wong; Jerome Kelleher; Gil McVean
Big Data Institute, University of Oxford, Oxford, UK

Inferring the ancestry of everyone

Inferring the evolutionary history of the genome is a fundamental problem in evolutionary biology. However, for sexual species, the genomic history, or ancestry, of individuals in a population is confounded by the fact that different regions of the genome have different histories. We have developed a technique («tsinfer») for inferring evolutionary trees from genetic variants at every point in the genome. The method scales to millions of individuals, providing comparable accuracy to full likelihood methods such as ARGweaver, and even outperforming them in cases such as selective sweeps. The method results in an «evolutionary encoding» for genetic variation data, allowing us to store genomes in a succinct format, suitable for rapid, genome-wide evolutionary analyses. In this talk I will briefly outline our evolutionary encoding technique and inference methodology, then present the patterns of deep ancestry revealed from the 1000 Genomes and the Simons Genome Diversity projects, as well as showing results from extending the analysis to include the million genomes from the UK Biobank. I will discuss current limitations of our approach, and our current focus on extending our ancestral inference to historical patterns in space and time.

Jonathan Mitchell[1,2]; Elizabeth Allman [2]; John Rhodes [2]
[1] Institut Pasteur, Paris, France; [2] University of Alaska Fairbanks, Fairbanks, USA

Testing n-Taxon Species Trees with the Multispecies Coalescent Model

Incomplete lineage sorting, where gene tree topologies can differ from species tree topologies, can be modeled by the multispecies coalescent model. Here we describe a test for an n-taxon species tree, with gene trees expected to arise in specific frequencies under the multispecies coalescent model. A substantial departure from these frequencies can be interpreted as evidence to reject the species tree and/or the multispecies coalescent model. A species tree may be rejected in favour of a network that models more complex biological processes such as hybridisation.

Donate Weghorn[1,2]; Felix Dietlein[3]; Eliezer M. Van Allen[3]; Shamil Sunyaev[4]
[1] Centre for Genomic Regulation (CRG), Barcelona, Spain; [2] Universitat Pompeu Fabra (UPF), Barcelona, Spain; [3] Dana-Farber Cancer Institute, Harvard Medical School, Boston, USA; [4] Harvard Medical School, Boston, USA

Probabilistic approaches to positive and negative selection inference on coding regions in cancer

Cancer genomics efforts have identified genes and regulatory elements driving cancer development and neoplastic progression. From an evolutionary perspective, these are subject to positive selection. Although elusive in current studies, genes whose wild-type coding sequences are needed for tumor growth are also of key interest. They are expected to experience negative selection and stay intact under pressure of incessant mutation. The detection of both significantly mutated (positive selection) and undermutated (negative selection) genes is completely confounded by the genomic heterogeneity of the cancer mutation rate.Here, I present two approaches we recently developed in order to address mutation rate heterogeneity to increase the power and accuracy of selection inference. Using a hierarchical model, we inferred the distribution of mutation rates across genes that underlies the observed distribution of the synonymous mutation count within a given cancer type. This enabled an inference of the posterior probability of nonsynonymous mutations under neutrality without additional parameters, however explicitly taking into account cancer type-specific mutational signatures, which are known to be highly distinct. We then extended the test for positive selection on genes through additionally integrating information at the single-nucleotide level, defining a «selection mutational signature». This test identifies genes with an excess of mutations in unusual nucleotide contexts, which deviate from the characteristic context around neutrally evolving passenger mutations.Application of the models to sequencing data from 28 cancer types demonstrates an increased power to detect known cancer driver genes. We discovered a long tail of novel candidate cancer genes with mutation frequencies as low as 1% and functional supporting evidence. The signal of negative selection is very subtle, but is detectable in several cancer types and in a pan-cancer data set. It is enriched in cell-essential genes identified in a CRISPR knockout screen, as well as in genes with reported roles in cancer.

Talks

Maryam Alamil[1]; Joseph Hughes[2]; Karine Berthier[3]; Cécile Desbiez[3]; Gaël Thébaud[4] and Samuel Soubeyrand[1]
[1] BioSP, INRA, 84914, Avignon, France; [2] MRC-University of Glasgow Centre for Virus Research, Glasgow, United Kingdom; [3] Pathologie Végétale, INRA, 84140 Montfavet, France; [4] BGPI, INRA, SupAgro, Cirad, Univ. Montpellier, Montpellier, France

A statistical learning approach to infer transmissions of infectious diseases from deep sequencing data

Pathogen sequence data have been exploited to infer who infected whom, by using empirical and model-based approaches. Most of these approaches exploit one pathogen sequence per infected host unit (e.g.,individual, household, field). However, data collected with deep sequencing techniques, providing a subsample of the pathogen variants at each sampling time, are expected to give more insight on epidemiological links than a single sequence per host unit. A mechanistic viewpoint to transmission and micro-evolution has generally been followed to infer epidemiological links from these data. Here, we investigate an alternative statistical learning approach for estimating epidemiological links, which consists of learning the structure of epidemiological links with a pseudo-evolutionary model and training data before inferring links for the whole data set. We designed the pseudo-evolutionary model as a semi-parametric regression function where the response variable is the set of sequences observed from a recipient host unit and the explanatory variable is the set of sequences observed from a putative source. We derived from this model a penalized pseudo-likelihood that is used for selecting who infected whom or who is closely related to whom, where the penalization is calibrated on training data. In order to assess the efficiency of the pseudo-evolutionary model and the associated inference approach for estimating epidemiological links, we applied it to simulated data generated with diverse sampling efforts, sequencing techniques (corresponding to diverse depths and read lengths), and stochastic models of viral evolution and transmission. Then, we applied it to three real epidemics: swine Influenza, Ebola and a potyvirus of wild salsify. Such an approach has the potential to be particularly valuable in the case of a risk of erroneous mechanistic assumptions and sequencing errors, it is sufficiently parsimonious to allow handling big data sets in the future, and it can be applied to very different contexts from animal, human and plant epidemiology.

Pedro Feijao[1]; **Cedric Chauve[2];** Leonid Chindelevitch[1]
[1] Schoolf of Computing Science, Simon Fraser University, Burnaby, Canada; [2] Department of Mathematics, Simon Fraser University, Burnaby, Canada

MLST genotyping of bacterial pathogens using whole-genome sequencing data

Multi-Locus Sequence Typing (MLST) is a genotyping method aimed at detecting which alleles of selected loci are present in a pathogen sample. Initially designed for MLST schemes considering a handful of loci (usually housekeeping genes), the method has recently been extended to core-genome and whole-genome schemes (cgMLST and wgMLST); in this context it has been shown that it provides genotyping results of quality at least as good as SNPs, and that it has good potential toward the detection of important phenotypes, such as antimicrobial resistance. However, traditional MLST algorithms do not scale well to large schemes that can contain hundreds to thousands of loci. In this talk we will describe progress toward a resource-efficient and accurate MLST tool, MentaLiST (https://github.com/WGS-TB/MentaLiST), developed in our group in the Genome Canada funded project PathoGiST. MentaLiST can handle very large MLST schemes with a small memory footprint while being extremely accurate especially for the detection of novel alleles, i.e. alleles present in a sample but not seen previously and not present in the initial MLST scheme.

Posters

Poster 1

Baltzis Athanasios Athanasios Baltzis[1,2]; Cedric Notredame[1,2]
[1] Centre for Genomic Regulation (CRG), Barcelona, Spain; [2] Universitat Pompeu Fabra (UPF), Barcelona, Spain

Reconstructing paralogues phylogenies using a concatenation approach

Although it is well established that gene duplication is a main driving force of evolution, inferring reliable phylogenetic relationships among paralogous genes remains a quite challenging process, as it is hampered by lots of confounding factors. Indeed, very ancient duplications, subsequent uneven gene losses and widespread heterotaxis all contribute towards preventing accurate tree reconstruction. It is, however, of great importance to build more accurate paralogous trees since such trees provide a unique insight into functional evolution, consequently, shedding light on the origin of basic cellular functions. In order to address this issue, we developed a novel method to reconstruct phylogenies of paralogous proteins based on the combination of information across related genomes. This method will be applied to several representative protein families containing an adequate number of paralogues and the produced trees will be evaluated using different techniques implementing also structural information.

Poster 2

Jutta Buschbom
Ahrensburg, Germany

When accuracy is of essence: The importance of sufficiently informative inference approaches for bridging the gap to conservation genetic applications

Forecasts of species' responses under quickly and potentially unpredictably changing climatic and environmental conditions, require reliable and detailed insight into the current state of a species, as well as, the processes shaping its responses to change. Such knowledge supports long-term sustainable conservation and management strategies and allows societies in dynamic contexts to come to adequate decisions and effective action in time.Sufficiently informative reference data and versatile inference approaches are at the core of conservation genetic tools. These are characterized by a need for high overall predictive accuracy and low predictive error rates in every-day application. Reference datasets spanning a species' evolutionary relationships and distribution range, its abiotic and biotic environments and its genomic record provide the necessary overall evolutionary and ecological context for applied tools. These tools, on the other hand, are versatilely optimized for concrete casework under operation conditions.Furthermore, in evolving and highly complex natural systems, hypothesis testing and event prediction require sufficiently informative, that is, information-rich, sequence-based genomic data. These support parameter-rich models and inference approaches, which are powerful enough to differentiate multiple processes that are simultaneously acting at different levels and might show continuous variation across space and time. Based on existing results for the genetic diversity of unlinked SSR markers in three species of white oaks (Quercus sect. Quercus) present in Central Europe, it is proposed that in addition to distribution-range-wide and genome-wide data, admixture linkage information ("ancestry blocks") is required to archive sufficiently accurate, decisive and reliable inference. The reconstruction of ancestry blocks likely is a necessary prerequisite for gaining useful and reliable insight into species membership, as well as, for arriving at population assignment to geographical region with sufficiently high spatial resolution for real-life tasks.

Poster 3

Markéta Harazim [1,2]; Lubomír Piálek [1,3]; Veronika Kováčová [4]; Jiří Pikula [4]; Jan Zukal [1,2]; Natália Martínková [1,5]

[1] Institute of Vertebrate Biology, Czech Academy of Sciences, Brno, Czech Republic; [2] Department of Botany and Zoology, Faculty of Science, Masaryk University, Brno, Czech Republic; [3] Department of Zoology, Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic; [4] Department of Ecology and Diseases of Game, Fish and Bees, University of Veterinary and Pharmaceutical Sciences Brno, Brno, Czech Republic; [5] Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic

Genome-wide associations in multiple bat species

Genome-wide association studies (GWAs) have become a popular tool to describe a relationship between phenotype and genotype in humans. With decreasing cost and increasing efficiency of next-generation sequencing methods, GWAs are available to ecological research. We investigated associations in non-model species and mapped traits present in multiple species. Sequencing methods, such as double-digest restriction site-associated DNA sequencing (ddRAD), screen for SNPs through randomly interspersed genome fragments in individuals or populations. We studied natural populations of hibernating bats that differ in susceptibility to disease, ability to tolerate disease or ability to deal with environmental stressors. The differences can be attributed to different hibernation conditions, previous pathogen exposure of individuals or historical pathogen exposure of populations. These could be associated with a certain genotypic profile. We used ddRAD sequencing data of 7 species of bats together with their phenotypic characteristics (e.g. capacity to carry a pathogenic fungus, susceptibility to disease) in a GWAs with a correction of individual phylogenetic distances constructed from the SNPs to avoid bias coming from the uneven distribution of variability in the tested dataset. We found weak associations of the studied phenotypic traits and the available SNPs.

Poster 4

Benjamin Nguyen-Van-Yen[1,2] ; Bernard Cazelles[2]

[1] Génomique fonctionnelle des maladies infectieuses, Institut Pasteur, Paris, France ; [2] Institut de biologie de l'École Normale Supérieure, Paris, France

General Purpose MCMC for Markov jump process estimation

When trying to estimate a complex stochastic model, the limited availabilityof data typically makes the observed data likelihood intractable [1].A common way to deal with that difficulty is to use "data-augmentation", that is,inferring the joint likelihood of the parameters of interestaugmented with latent variables to make it tractable.This has proven useful in various settings [2,3] and maybe most notably,this strategy has been very successful for phylogeny estimation,like it is done in BEAST [4].But an important challenge with data augmentation is the high dimensionof the resulting state space, which can lead to slow convergence and mixing.All the subtlety in the approach then lies in designing a suitable proposal for thelatent state. Each particular problem then requires a carefully crafted proposal.For instance, BEAST implements many different kinds of tree moves, that arechosen alternatively to explore tree space effectively.To apply data augmentation more easily to different models, a lot of efforthas gone into finding more general data augmentation schemes andproposals [3]. Building on top of this, we propose a data augmentation scheme-

that is applicable to any continuous time Markov jump process.We use an equivalent formulation of a jump process as the solution ofa stochastic integral equation with respect to a Poisson random measure.The Poisson random measure can be simulated first, independently of the realizationof the model, then the system is integrated from those points in a deterministic manner.This can be done with exact and approximate algorithms analogous to classical SSA and tau-leaping.We can then use this discrete measure as a latent variable, and at eachiteration we explore the discrete measure space by proposing to add and removemany points at once.We show that this is an effective strategy for inferring large epidemics from coarse-grained data.We also propose possibilities to adapt this method to more rich data like sequence data, with a different integration method and proposal,even though it becomes much more challenging to obtain good mixing.We believe this Poisson random measure formulation to be a promising avenuefor stochastic model estimation.
[1] Philip D O'Neill. "A tutorial introduction to Bayesian inference for stochas-tic epidemic models using Markov chain Monte Carlo methods". In: Math-ematical biosciences 180.1-2 (2002), pp. 103–114.[2] Simon Cauchemez et al. "A Bayesian MCMC approach to study transmis-sion of influenza: application to household longitudinal data". In: Statisticsin medicine 23.22 (2004), pp. 3469–3487.[3] Jonathan Fintzi et al. "Efficient data augmentation for fitting stochasticepidemic models to prevalence data". In: Journal of Computational andGraphical Statistics 26.4 (2017), pp. 918–929.[4] Alexei J Drummond and Andrew Rambaut. "BEAST: Bayesian evolution-ary analysis by sampling trees". In: BMC evolutionary biology 7.1 (2007),p. 214.

Poster 5

Cyriel Paris[1], Simon Boitard[1], Bertrand Servin[1]
[1] Institut national de la recherche agronomique (INRA), Toulouse, France

Detecting Selection from Genomic Time Series with an Efficient Continuous Approximation of the Wright-Fisher Process

Detecting genomic regions under selection is one of the main issues in population genetics. While most methods exploit different kinds of patterns observed in present time data, recent DNA sequencing techniques allow to get more and more time series genomic data. A common modeling approach of these data is to describe the temporal evolution of an allele frequency as a Markov chain. Based on this principle, several methods have been proposed to infer selection intensity for a given polymorphism. One of the main differences between these methods lies in how they model the Markov chain's transition probabilities. Indeed, although the Wright-Fisher model is a natural choice, its computational cost is prohibitive for large population sizes. To overcome this limitation, other models consider approximations to the Wright-Fisher model, based on continuous transition densities, possibly allowing for fixations. Using simulations, we compared the performance of several of these approximations with respect to their computational time, power to detect selection and estimation of the selection coefficient. To this aim, we developped a new generic Hidden Markov Model likelihood calculator and applied it on simulations of various scenarii in terms of population sizes, selection intensities and data collection times. We show that the Beta with spike approximation, which accounts for fixation probabilities, provides a very good approximation to the Wright-Fisher process for a computational cost that does not increase with population size.

Poster 6

Pijus Simonaitis [1], Annie Chateau [1,2,3], Krister Swenson [1,2,3]
[1] LIRMM, Université de Montpellier, Montpellier, France; [2] Institut de Biologie Computationnelle (IBC), Montpellier, France; [3] CNRS

Weighted Genome Rearrangements

We develop mathematical models and algorithms that allow us to relate chromatin conformation to evolutionary scenarios of genome rearrangements in eukaryotes. The benefit of this is two-fold. First, we can make inferences about the evolution of the chromosome conformation. Second, we can put a confidence measure on ancestral rearrangements and ancestral breakpoints using chromatin conformation.Our work is motivated by a couple of hypotheses. First, that the sequences undergoing rearrangement need to be in close spatial proximity in the nucleus to become joined [1]. And second, that genome's spatial organization is somewhat conserved across evolutionary distances [2].The advent of high-throughput technologies like Hi-C and ChIA-PET together with an increasing number of fully assembled genomes provide a timely opportunity to study the breakpoints of the rearrangements in detail.From a computational perspective, our goal is to define a framework for cost-constrained genome rearrangements and to devise algorithms for finding chromosome rearrangement scenarios within this framework [3]. Using Hi-C data we can infer the evolutionary scenarios maximizing the co-locality of the breakpoints. This enables us to study our hypotheses in detail and our preliminary results concerning Drosophila species are in line with them. This paves the road for the future study of more complex rearrangement histories in mammals. [1] - Triggers for Genomic Rearrangements: Insights into Genomic, Cellular and Environmental Influences, Mani and Chinnaiyan, Nature Reviews Genetics, 2010.[2] - Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture, Rudan et. al., Cell Reports, 2015.[3] - A General Framework for Genome Rearrangement with Biological Constraints, Simonaitis et. al., RECOMB-CG, 2018.

Poster 7

ZHUKOVA Anna, Institut Pasteur & CNRS, Paris, France

PastML: Fast inference and visualization of ancestral scenarios

Poster 8

Frédéric Lemoine[1,2] , Vincent Lefort[3], Fabien Mareuil[2], Sarah Cohen-Boulakia[4] , and Olivier Gascuel [1,3]
[1] Unité Bioinformatique Evolutive, C3BI USR 3756, Institut Pasteur & CNRS, Paris, France; [2] Hub Bioinformatique et Biostatistique, C3BI USR 3756, Institut Pasteur & CNRS, Paris, France; [3] Méthodes et Algorithmes pour la Bioinformatique, LIRMM UMR 5506, Université de Montpellier & CNRS, Montpellier, France. [4] Laboratoire de Recherche en Informatique, Université Paris-Sud, CNRS UMR 8623, Université Paris-Saclay, Orsay, France.

NGPhylogeny.fr: New Generation Phylogenetic Services for Non-Specialists

Phylogeny.fr, created in 2008, has been designed to facilitate the execution of phylogenetic workflows, and is nowadays widely used. However, since its development, user needs have evolved, new tools and workflows have been published, and the number of jobs has increased dramatically, thus promoting new practices, which motivated its refactoring. We developed NGPhylogeny.fr to be

more flexible in terms of tools and workflows, easily installable, and more scalable. It integrates numerous tools in their latest version (e.g. TNT, FastME, MrBayes, etc.) as well as new ones designed in the last ten years (e.g. PhyML-SMS, FastTree, Noisy, BOOSTER, etc.). These tools cover a large range of usage (sequence searching, multiple sequence alignment, model selection, tree inference and tree drawing) and a large panel of standard methods (distance, parsimony, maximum likelihood and bayesian). They are integrated in workflows, which have been already configured ("One click"), can be customized ("Advanced"), or are built from scratch ("A la carte"). Workflows are managed and run by an underlying Galaxy workflow system, which makes workflows more scalable in terms of number of jobs and size of data. NGPhylogeny.fr is deployable on any server or personal computer. Thanks to its architecture, made of a python/django web interface and a galaxy workflow system, NGPhylogeny is:- expressive, offering a very large panel of phylogenetic tools, - flexible and modular, allowing to easily add or remove tools, - scalable, able to support large-scale analyses (up to 10,000 sequences), - turnkey, avoiding user to do not install anything on their own computers, - user-adaptable, providing a large panorama of usages from pure end-users to bioinformaticians with technical skills.NGPhylogeny.fr is freely accessible at https://ngphylogeny.fr

Poster 9

Carlos Santana-Molina[1]; Elena Rivas-Marin[1]; Ana Rojas[1,2]; Damien P. Devos[1]
[1] Centro Andaluz de Biología del Desarrollo (CABD), Seville, Spain; [2] Instituto de Biomedicina de Sevilla (IBIS), Seville, Spain;

Origin and evolution of polycyclic-triterpenes synthesis

The terpene family contains carotenoids, hopanoids and sterols, all of which are biosynthetically related. While carotenoids are found in the three domains of life, hopanoids are mostly found in bacteria and sterol are limited to eukaryotes with few, but growing, bacterial exceptions (Wei et al. 2016). Due to their important role in eukaryotes, the combination of their omnipresence in all eukaryotes with their dearth in prokaryotes had previously suggested that sterol synthesis was an eukaryotic innovation (Cavalier-Smith, 2002). Thus, the presence of sterol in a few bacteria was deemed to be the results of lateral gene transfer from eukaryotes (Desmond and Gribaldo, 2009). To elucidate the origin and evolution of polycyclic triterpene synthesis pathways is important for various reasons, such as their role in eukaryogenesis or their importance as biomarkers in fossil records for geobiology. In contrast to previous analyses that mostly focused on the cyclases, here we revisited the phylogenies of the main enzymes involved in triterpene synthesis. Combined to gene neighborhood analysis and phylogenetic profiling, our results suggest that HpnCDE is the ancestral squalene synthesis pathway and might be more metabolically versatile. Sqs, on the other side, appear to be more specialized leading to an individualization of synthesis pathways. According to this and to our reconstruction, both squalene pathways pressume to have bacterial origin. It is likely that the hopanoid cyclase, SHC, was already present in the ancestor of bacteria as previously suggested (Firckey and Kannenberg, 2009). No triterpene cyclases are observed in Archaea, which together with our phylogenetic reconstructions, argues against the eukaryotic origin of sterol synthesis and challenges its value as an "eukaryotic specific" hallmark.In parallel to these computational analyses, we have also carried out experimental analyses regarding to the sterol synthesis in a very striking bacteria, the Planctomycete Gemmata obscuriglobus. This bacteria have a complex endomembrane-system with unknown function so far (Santarella-Mellwig et al.

2010) and its sterol genes are far related to the eukaryotic ones. Indeed, when sterol genes are interrupted in G. obscuriglobus, we observe membrane phenotypes, cell division defects and growth deficiency leading to death, suggesting that sterol could play an essential role in this bacteria, a fact previously not reported in other bacteria.Considering the current most accepted view of tree of life (Eme et al. 2017), altogether supports the idea that sterol was originated in Bacteria domain and thus, ancestral eukaryotes gained sterol genes from a bacterial contribution which probably enhanced basal eukaryotic biological process such as the phagocytosis.

Poster 10

Thimothée Virgoulay[1][2]; Raphaël Leblois[2]; François Rousset[1]; François-David Collin[3]; Jean-Michel Marin[3];

[1]Institut des Sciences de l'Evolution de Montpellier (ISEM), Montpellier, France; [2] Centre Biologique pour la Gestion des Populations (CBGP), Montferriez-sur-Lez, Montpellier; [3] Institut Montpelliérain Alexander Grothendieck, Montpellier, France

Inférences démographiques et historiques à partir de données génomiques sous des modèles spatialisés réalistes : vers une prise en compte du paysage

L'analyse du polymorphisme génétique neutre permet d'estimer des paramètres démographiques et historiques des populations tels que des tailles ou des densités de population, des paramètres de dispersion, des temps de divergence ou des changements démographiques passés. Ces analyses reposent sur la combinaison (1) de modèles stochastiques de l'évolution des populations tels que le coalescent de Kingman (1982) pour des locus indépendants ou le graphe ancestral de recombinaison (Hudson 1983, Griffiths et Marjoram 1997) prenant en compte la recombinaison entre séquences ; et (2) des méthodes d'inférence statistique, dont les plus puissantes sont basées sur l'estimation de la vraisemblance pour les modèles d'évolution les plus simples (Kuhner 2009, Rousset et al. 2018), ou sur la comparaison de simulations avec les jeux de données réels (à travers un ensemble de statistiques résumées) pour les modèles plus complexes (méthodes « Approximate Bayesian Computation » ABC, Beaumont 2010, Marin et al. 2012). L'enjeu est de développer et tester des outils inférentiels adaptés à une classe bien spécifique de modèles stochastiques de génétique des populations : les modèles démographiques spatialisés. En effet, chez de nombreuses espèces, la dispersion des individus est limitée dans l'espace: les individus se reproduisent préférentiellement avec des individus proches géographiquement. Les modèles spatialisés d'isolement par la distance (IBD) en habitat continu prennent en compte ces caractéristiques, et permettent notamment d'estimer certaines caractéristiques de dispersion et de densité des populations. Cependant le développement de nouvelles méthodes d'analyses spatialisées reste relativement limité, du fait certainement de la lourdeur de mise en œuvre des méthodes d'inférence sur des données démo-génétiques spatialisées. Les principales méthodes d'inférence existantes sont encore basées sur l'utilisation des F-statistiques, et permettent uniquement l'estimation de la taille de voisinage, le produit de la densité par la dispersion (Rousset 1997, 2000). Une méthode d'inférence par maximum de vraisemblance, et utilisant donc toute l'information des données génétiques, a été développée plus récemment ainsi que la mise au point de nouvelles méthodes d'inférences basées sur la simulation (« Approximate Bayesian Computation using Random Forest », ABC-RF, Pudlo et al. 2015, Marin et al. 2017 ; ou « the summary-likelihood method », SL, implémentée dans le package R Infusion, Rousset 2016).Ces deux avancées majeures permettent aujourd'hui de considérer des

modèles spatialisés réalistes pour lesquels la simulation est relativement lente, ainsi qu'un très grand nombre de marqueurs, afin d'inférer avec plus de détails et de précision le fonctionnement démographiques des populations dans l'espace et dans le temps que ce qui est permis avec les méthodes actuelles.

Poster 11

Armando Arredondo[1]; Olivier Mazet[1]; Lounès Chikhi[2]; Willy Rodriguez[1]
[1] Institut National des Sciences Appliquées (INSA), Toulouse, France; [2] Centre National de la Recherche Scientifique (CNRS), Toulouse, France

Introducing SNIF: An inferential framework for structured demographic inference

In the context of demographic inference, one of the characterizing signals that can be obtained from genomic sequences is the IICR, or Inverse Instantaneous Coalescence Rate. For a single large and isolated population with no intrinsic structure this function matches the size of the population, which can be partially recovered by methods like PSMC. But we also know that the presence of demographic structure has a strong effect on the IICR. We model this population structure as a continuous-time Markov process, where the states of the process represent the different ways in which a small number of genetic samples from the present population migrate between sub-populations and coalesce with each other. One of the main aspects of this work is to explore the distribution of these random variables under different parameters of the structure. This involves finding efficient ways to compute the probability distribution, density functions and expected values of coalescence times. Computationally speaking, this usually involves the multiplication and exponentiation of relatively large matrices. Having a fast and reliable method of computing the IICR curves enables the use of meta-heuristics in order to perform curve fitting, and thus, inference of selected demographic parameters within a pre-specified parametric space.

Poster 12

Raphaël Mourad
Université Paul Sabatier, Toulouse, France

Studying 3D genome evolution using genomic sequence

BackgroundChromosomes are tightly packed in three dimensions (3D) such that a 2-meter long human genome can fit into a nucleus of approximately 10 microns in diameter. Over the past years, the 3D chromosome structure has been comprehensively explored by chromosome conformation capture combined with high-throughput sequencing technique (Hi-C) at an unprecedented resolution [1]. The 3D genome is essential to numerous key processes such as the regulation of gene expression and the replication-timing program. In vertebrates, chromatin looping is often mediated by CTCF, and marked by CTCF motif pairs in convergent orientation [1]. Comparative Hi-C revealed that chromatin looping evolves across species [2-4]. However, Hi-C experiments are complex and costly, which currently limits their use for evolutive studies over a large number of species. MethodHere, we propose a novel approach to study for the first time the 3D genome evolution in vertebrates using the genomic sequence only, e.g. without the need for Hi-C data. Therefore, this approach allows a comprehensive analysis of vertebrate 3D genomes whose number is exponentially increasing due to ungoing large sequencing projects such as the Vertebrate Genomes Project (VGP). The approach is simple and relies on comparing the distances

between convergent and divergent CTCF motifs along the genome (ratio R). Such motifs can be called from vertebrate genome assemblies to compute R. R can then be used to assess the strength of CTCF-mediated looping in a genome and also to identify significant differences between species. R can also be used for phylogenetic analyses of CTCF-mediated looping and for ancestral R reconstruction in ancestral genomes. ResultsWe show that R is a powerful statistic to detect CTCF looping encoded in the human genome sequence, thus reflecting strong evolutionary constraints encoded in DNA and associated with the 3D genome organization. When comparing R across vertebrates, our results reveal that the distance between convergent motifs which underly CTCF looping and TAD organization evolves over time and suggest that ancestral character reconstruction can be used to infer R in ancestral genomes. References[1] Suhas S. P. Rao, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159(7):1665-1680, December 2014.[2] Carlos Gomez-Marin, et al. Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. Proceedings of the National Academy of Sciences, 112(24):7542-7547, June 2015.[3] Peter Heger, et al. The chromatin insulator CTCF and the emergence of metazoan diversity. Proceedings of the National Academy of Sciences, 109(43):17507-17512, 2012.[4] Matteo Vietri-Rudan, et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Reports, 10(8):1297-1309, March 2015.

Poster 13

Marie Morel[1,2]; Frédéric Lemoine[1]; Olivier Gascuel[1,3]

[1] Unité de Bioinformatique Evolutive, Institut Pasteur, C3BI USR 37 56 IP CNRS, Paris, France; [2] Centre de Recherche Interdisciplinaire, Paris Diderot Université, Paris, France [3] Méthodes et Algorithmes pour la Bioinformatique, LIRMM UMR 5506, Université de Montpellier & CNRS, Montpellier, France.

Revealing Convergent Substitutions without knowledge of phenotype

Convergent evolution at the molecular level is the independent acquisition of identical DNA or protein substitutions in different lineages. Molecular convergence is a more and more studied phenomenon thanks to the increasing amount of genome wide data. Indeed, the acquisition of similar traits at the phenotypic level has been studied for many years, but without really being able to explain it at the genetic level[1]. Several studies now focus on understanding to what extent phenotypic convergence can be related to similar changes at the genetic level[2,3].In the case of higher eukaryotes, convergent evolution at the genetic level is assumed to be quite rare, but in organisms with very high mutation rate such as viruses, we find several examples of the independent surge of the same mutations in different lineages[4,5]. These convergent mutations are important to study since they can lead to 1) the prediction of evolutionary pathways for viruses under selective constraints, 2) better understand these constraints, for example in the case of treatments or host specificity, and 3) evaluate the interest of using these mutated regions as targets for therapeutic drugs. However, for viruses, information on the phenotype can be difficult to retrieve and thus it is complicated to apply existing methods for the search of convergent substitutions.Here we try to answer the following question: Can we detect convergent mutations in viral alignments without any knowledge on the phenotype? To do so, we developed a method based on simulations and ancestral states reconstruction to compare the number of independent apparitions of a muta-

tion with what we could have expected under a neutral model of evolution. We tested it on real datasets with known convergent mutations (drug resistance mutations in HIV) and simulations for which we introduced positions under convergence. The results show that known (real, simulated) convergent substitutions are well detected, but the method produces a substantial fraction of false positives when model violation occur. This method is compared with dn/ds approaches and we assess its performance in measuring the "convergenceness" at the gene level.[1] Storz, J.F. (2016). Causes of molecular convergence and parallelism in protein evolution. Nat Rev Genet 17, 239–250.[2] Rosenblum, E.B., Parent, C.E., and Brandt, E.E. (2014). The Molecular Basis of Phenotypic Convergence. Annual Review of Ecology, Evolution, and Systematics 45, 203–226.[3] Rey, C., Gueguen, L., Semon, M. and Boussau, B. (2018). Accurate Detection of Convergent Amino-Acid Evolution with PCOC. Molecular Biology and Evolution 35.9, 2296–2306,[4] Bertels, F., Metzner, K.J., and Regoes, R.R. (2017). Convergent evolution as an indicator for selection during acute HIV-1 infection. BioRxiv 168260.[5] Vignuzzi, M., and Higgs, S. (2017). The Bridges and Blockades to Evolutionary Convergence on the Road to Predicting Chikungunya Virus Evolution. Annu Rev Virol 4, 181–200.

Poster 14

Nathanael Zweig [1], Alexandre Baldachino [1], Bianca H. Habermann [1]
[1] Aix-Marseille University, CNRS, IBDM UMR7288, Marseille, France;

morFeus and beyond: searching orthologs in the twilight zone of sequence similarity in low and high throughput

Identifying orthologous proteins is one of the key tasks in computational biology: we need to know a protein's conservation across species to understand its evolution. Orthologs also tell us, whether the process a protein is involved in, is conserved beyond model species and across kingdoms. The level of sequence conservation between orthologs can however be below the detection limit of standard software and settings in the so-called twilight zone of sequence similarity. Such remote orthologs are therefore difficult to detect using standard approaches.We have developed web-based method, morFeus, for identifying remote orthologs in low throughput (Wagner, et al., 2014, PMID: 25096057). By combining clustering of BLAST alignments, iterative reciprocal BLAST searches for reciprocal best hits, as well as network scoring, we are able to detect orthologs in the in the twilight zone of sequence similarity.Based on novel developments in sequence similarity searching (using Hidden Markov Model (HH-) comparisons (Meier and Söding, 2015, PMID: 26496371)) and novel observations (Habermann, 2016, Evolutionary Biology pp 393-419 (Springer)), we are taking the detection of remote orthologs on genome-scale and have developed a stand-alone package for proteome-wide searches for remote orthology (RemOtF). A GUI-based web-interface for analyzing RemOtF results makes interpretation of proteome-wide data user-friendly and easy.

Poster 15

Tetyana Nosenko, Jörg-Peter Schnitzler
Research Unit Environmental Simulations, Helmholtz Center of Munich, Germany

Incomplete genome annotation vs. missing function: The case of plant Antimicrobial Peptides.

Comparative genomics relies pretty much on quality of genome sequence assemblies and genome annotations. While the assembly of genome improved tremendously with the recent development of sequencing technologies, the genome annotation completeness and accuracy vary significantly depending on gene prediction algorithms and parameter thresholds used. The different quality of genome annotations can be erroneously translated into the between-species functional difference.Gene prediction methods are usually based on sequence similarity to gene models previously annotated for closely related species. Often, evidence of expression (EST or RNA-Seq data) is a prerequisite for including a model in the high-confidence gene set. These comparative and transcript-based approaches result in high confidence predictions for conserved genes with high-to-moderate expression levels. However, genes characterized by low sequence conservation and low-level conditional expression have a high chance to be overlooked or filtered away by automated genome annotation.In our study we focus on genes encoding plant antimicrobial peptides (AMPs). AMP genes are short, rapidly evolving, and often expressed only in response to specific stress-factors. These features make annotation of the AMP-encoding genes challenging. For example, over 300 genes encoding defensin-like proteins (DEFLs; a sub-group of AMPs) have been identified in Arabidopsis [1], while official annotations of recently released genomes for three Fagaceae species, Fagus sylvatica, Quercus suber, and Q. robur, contain only eight, four, and none DEFLs, respectively.To identify the AMP-encoding loci in the Fagaceae genomes, we applied two complementary approaches: (1) screening de-novo transcriptome assemblies with the key-amino acid patterns and (2) combining ab initio gene model prediction with profiles of conserved protein blocks. Both the key-amino acid patterns and protein block profiles were inferred from multiple sequence alignments of known plant AMPs. Additional Fagaceae-specific protein profiles were constructed using alignments of AMP transcripts detected at the first step of analysis. Q. robur expression data for different tissues and conditions were obtained from publicly accessible databases and used to verify structure, functionality, and expression specificity of the predicted AMP genes.Our study demonstrates that the integration of the profile-based methods into the standard gene prediction workflow is necessary for increasing genome annotation accuracy.Silverstein et al. Plant Physiology, 2005, 138: 600-610.

Poster 16

Leila Mansouri[1]; Cedrik Magis[1]; Cedric Notredame[1,2]

[1] Centre for Genomic Regulation (CRG), Barcelona, Spain; [2] Universitat Pompeu Fabra (UPF), Barcelona, Spain

Systematic use of structural information to exclude low reliability sequences in large-scale multiple sequence alignments

Multiple Sequence Alignments (MSA) are routinely used for many essential biological applications such as evolutionary modelling or structural prediction. Unexpectedly, the rapidly increasing number of available sequences has raised serious issues for at least two main reasons: firstly of all, methods able to align the data have been trained and tested on benchmarking dataset that are not able to mimic the extent of information and diversity present in large-scale alignments inputs[1]; secondly because the accuracy of the alignments decreases when aligning more than 1000 sequence [2]. Assessing large scale alignment accuracy is therefore a goal just as desirable as computing these large-scale models. We have therefore developed a methodology allowing the

systematic assessment of the quality and reliability of large-scale multiple sequence alignments. The aim of this methodology is not only to assign an accuracy score but also to poorly aligned sequences that could compromise down-stream analysis.Our methodology relies on the idea that a small number of sequences with known structures embedded within a larger dataset of homologues can be used to infer the structural correctness of the alignment of neighbour sequences without known structures. In order to test this hypothesis, we generated a dataset applying an automated pipeline developed in Nextflow, the BenchFam pipeline, on the Pfam database (release 28.0). The resulting dataset consists of 666 Pfam families each containing 10 or more structure-endowed sequences deposited in the PDB; it contains of 32,275,933 sequence, out of which 15,248 are structure-endowed. We developed a method to assess the accuracy of the alignment of each sequence in the large-scale alignment testing the assumption that the accuracy of a given sequence could be predicted by looking at the accuracy of the reference sequences building up its neighbourhood. The method is based on the k-Nearest Neighbour algorithm, implemented as a regressor. The idea behind this machine learning algorithm is to identify the k- nearest neighbours of the target Y and use their distance from Y as a way to weight their contribution to the prediction.In our implementation, we selected the 8-nearest neighbours and used them to predict the accuracy of their nearest neighbour. The use of a similarity threshold makes it possible to exclude the neighbours considered to be too distant to any structure-endowed sequence. Our benchmarks indicate that when using a 95% threshold on the pearson correlation between the predicted accuracy score and the one estimated using the reference structures, a prediction can be made on 66.26% of the sequences.[1]Linder, C. R., et al (2010). PLoS Currents, 2, RRN1195. [2]Sievers, F., et al (2011Molecular Systems Biology, 7(539).

Poster 17

Willy Rodríguez[1]; Olivier Mazet[1]; Simona Grusea[1]; Armando Arredondo[1]; Josué M. Corujo[2]; Simon Boitard[3]; Lounès Chikhi[4,5]

[1] Institut de Mathématiques de Toulouse, Université de Toulouse, Institut National des Sciences Appliquées, Toulouse, France; [2] Facultad de Matemática y Computación, Universidad de La Habana, La Havana, Cuba; [3] GenPhySE, Université de Toulouse, INRA, INPT, INP-ENVT, Castanet Tolosan, France; [4] Laboratoire Évolution & Diversité Biologique (EDB UMR 5174), Université de Toulouse Midi-Pyrénées, CNRS, IRD, UPS, Toulouse, France; [5] Instituto Gulbenkian de Ciência, Oeiras, Portugal.

The Non Stationary Structured Coalescent (NSSC): modeling demographic changes on structured populations

The increasing amount of genomic data currently available is expanding the horizons of population genetics. Within the last ten years, a wide range of methods allowing to reconstruct past population size changes from genome-wide data have been developed. At the same time, there has been an increasing recognition that population structure can generate genetic data similar to those generated under models of population size change. Some works have addressed the question of distinguishing patterns corresponding to a structured population from those coming from a population with changes in size. However, there is a shortage of more realistic models allowing to incorporate demographic changes into a population that is structured. In this talk I will present the NSSC framework, an extension of the Structured Coalescent (Herbots 1994) that incorporates

past demographic events into models of population structure. The NSSC takes advantage of the Markov property and uses the matrix exponential (Hobolth et al. 2011) to compute the distribution of coalescence times under complex structured models. Finally, I will illustrate how this framework can offer a different perspective to explain the IICR (Mazet et al. 2016) obtained from human and Neanderthal genomes.

Poster 18

Rui Borges[1]; Carolin Kosiol[1,2]
[1] Institute of Population Genetics, Vetmeduni Vienna, Vienna, Austria; [2] Centre for Biological Diversity, University of St Andrews, St Andrews, United Kingdom

Polymorphism-aware phylogenetic models provide a consistent estimator of the species tree

In phylogenetic theory, a consistent estimator is an estimator having the property that as the number of sites of the sequence alignment increases indefinitely, the resulting series of estimates converge in probability to the true phylogenetic tree. Simple phylogeny estimation methods using standard substitution models (e.g., maximum likelihood or Bayesian inference under the Jukes-Cantor model) enjoy statistical consistency, but the same principle cannot be extended to more complex and general methods of tree inference. The polymorphism-aware phylogenetic models (PoMo) is such an example of an alternative approach for tree estimation accounting for incomplete lineage sorting. While PoMo can be more broadly classified as a nucleotide substitution model, PoMo adds a new layer of complexity by accounting for the population-level evolutionary processes (such as mutations, genetic drift, and selection) to describe the evolutionary process. To do so, PoMo expands the standard substitution models to include polymorphic states, thereby, permitting to account for multi-individual data. PoMo has received substantial attention from the evolutionary community, and several publications have employed it to solve a wide range of evolutionary questions: e.g., disentangling phylogenetic relationships among baboon species, describing the phylogeographic history of great apes and estimating patterns of GC-bias and mutational biases from population data. Recently, PoMo was integrated into a Bayesian inferential framework. All this raised the question of whether PoMo is a statistically consistent phylogeny estimator for standard phylogenetic datasets. Building upon the formal results provided by Steel (JTB, 2013) and properties of the PoMo rate matrix and stationary distribution, we present a proof that the maximum posterior tree is a statistically consistent estimate of the true evolutionary tree. This result shows that PoMo is a statistically sound method of phylogeny inference, and it reassures further investigations and use of PoMo methods on real datasets.

Poster 19

Presented jointly by Andreas Futschik and Marta Pelizzola (further contributors: Merle Behr, Housen Li) JKU and University of Veterinary Medicine

An approach to reconstruct haplotypes from pool sequencing data

Poster 20

Marta Pelizzola [1,2], Andreas Futschik [2,3], Merle Behr [4], Housen Li [5]
[1] Institut Für Populationsgenetik, Veterinary Medicine University, Vienna, Austria, [2] Vien-

na Graduate School of Population Genetics, Veterinary Medicine University, Vienna, Austria, [3] Department of Applied Statistics, Johannes Kepler University, Linz, Austria, [4] Department of Statistics, University of California Berkeley, 367 Evans Hall, Berkeley, CA 94720, [5] Institute for Mathematical Stochastics, University of Göttingen, Goldschmidtstraße 7, 37077 Göttingen

Experimental evolution with haplotype data: How and why to reconstruct haplotypes from pool sequencing data

Evolve and Resequence (E&R) methods follow many replicate populations of a given model organism, like C. elegans or Drosophila, throughout time in order to understand how it adapts to different conditions. It has been shown that performing pool sequencing of one population instead of individual sequencing can give the same sequencing coverage with much lower costs. Furthermore, the power of detecting SNPs being targets of selection stays high. However, by doing so, the haplotype information is lost. Starting from pool sequenced data, our goal is to characterise the haplotype structure and to disentangle the possible patterns of selection in the data. We are investigating trajectories of haplotypes given different selection pressures, how the haplotype structure can help in identifying the truly causal SNPs and which starting condition of the experiment would favour the haplotype reconstruction and the detection of the SNPs driving adaptation.The first step of our research consists in adapting a work that exploits minimax theory to estimate the design matrix and the regression coefficients when only the response is given (Behr and Munk [2017]) and using it to reconstruct the haplotype structure and the haplotype frequency given the allele frequency from pool sequencing only. Afterwards, we use the information in the trajectories of haplotypes with the allele frequency trajectories to test for significant allele frequency changes. We currently work on design guidelines that help identifyingSNPs relevant for local adaptation in E&R experiments.ReferencesBehr, M. and Munk, A. (2017). Minimax estimation in linear models with unknown finite alphabet designThis application is for a joint presentation of Marta Pelizzola and Andreas Futschik

Poster 21

Alfried P. Vogler[1][2]; Thomas J. Creedy[1]

[1] Natural History Museum and Imperial College London, London, UK; [2]Institute Pasteur, Paris, France

Site-based studies of biodiversity to assemble the Tree-of-Life

As sequence information becomes available for an ever greater number of species, we will be able to assemble the tree-of-life to a high degree of completion, ultimately at the species-level. However, currently the data are too inconsistent and methods for tree construction not sufficiently scalable. In addition, the sampling of species for DNA analyses is challenging, given that we don't even know how many species there are on Earth, possibly not even to an order of magnitude. Current approaches to building the Tree rely on taxonomic sampling and data mining from public databases, but thave resulted in haphazard species representation and inconsistent gene choice. Here we propose a 'site-based' approach based on in-depth sampling of specific localities of species richness that are expected to harbour a large proportion of the extant global clade diversity. The sampling targets highly species rich arthropods obtained with passive traps. In a pilot study

we use the Coleoptera (beetles), the arguably largest radiation of animals, as a model for the site-based sampling. The samples are subjected to shotgun sequencing of bulk specimens, and the metagenomic mixture is then used to assemble mitochondrial genomes for each of the species present. In addition, short (meta)barcode sequences are collected for a greater number of species. The approach is now rolled out on a larger scale in the SITE-100 project that generates uniform sequence data from 'genomic observatories' from strategically chosen sites around the globe. Here, we present initial results from this study using a tree of ~5000 mitogenomes and ~14000 species represented by barcodes from anonymous environmental sampling at high-biodiversity sites from all continents. We assess the quality of the trees, and test the phylogenetic information content of barcode sequences when combined with the full-length mitogenomes. The study shows that barcodes are sufficient if mitogenomes of close relatives are present. We also find that phylogenies from site-based studies show a high degree of geographic structure, revealing subclades of various sizes that are restricted to major biogeographic regions of the world. Thus, the degree of geographic structure can be used to evaluate the quality of tree estimates for very largely trees. Within this framework each new sequence can immediately be placed in the context of the species' historical biogeography. In a next step these site-based studies are linked to taxonomy-based data bases (Genbank and Barcode-of-Life), whereby mitochondrial genomes play a critical role as a scaffold that links different data types (from full nuclear genomes to short barcodes). New methods are needed to generate trees at the scale of hundreds of thousands of species.

Poster 22

Luc Blassel [1]; Anna Zhukova [1]; Olivier Gascuel[1]

[1]Unité Bioinformatique Evolutive, Institut Pasteur, C3BI USR 3756 IP & CNRS, Paris, France

Machine learning approaches to detect resistances in HIV

Drug resistance mutations (DRMs) appear in HIV when under treatment pressure. These DRMs limit treatment options at a population level, due to increased transmission of resistant strains in the treatment-naive population [1]. Expert lists of surveillance DRMs (SDRMs) have been established [2], but they are not exhaustive, as treatment failure can happen without observing any known DRMs.In this study we aim to find new potential DRMs and bring to light poorly studied epistasis effects. To do this we did not use the classical testing approach [3], but instead used machine learning techniques, to avoid the loss of statistical power inherent to multiple testing. We have a European dataset of approximately 55,000 HIV-1 pol RT sequences annotated with patient treatment status. It is accepted that successful treatment results in an undetectable viral load, therefore treated samples were considered as resistant and naive ones as non-resistant. Amino-acid values at each position were used as categorical features and "one-hot" encoded. Features corresponding to known DRMs were removed. Naive and treated classes were balanced, and the whole dataset was split into training (80%) and testing (20%) sets. Interpretable classifiers (logistic regression and random forest) were trained on a binary classification task to discriminate naive and treated sequences, separately on each subtype to eliminate potential confounding effects. Several interpretation measures were used (weights, feature importance and feature contribution [4]) to determine the features on which the classifiers base their decisions. The most important

ones for both classifiers were selected for closer study. Epistasis effects were also measured, using these classifiers, by incrementally restricting the number of features being used, either by regularization in the case of logistic regression, or by limiting the depth of the decision trees with random forest.Even after the removal of all known DRMs from the training set, the trained classifiers have accuracies of 53% and 56% measured by 5-fold cross-validation. This shows the presence of some signal in the data potentially due to unknown DRMs and episasis effects. Furthermore, we tested the difference in prevalence in naive and treated samples for the 10 most important features of our classifiers on a separate, unseen African dataset. For 6 of these the prevalence was significantly higher in the treated population. Further in-vitro analysis will be needed to verify if they are indeed DRMs.[1] Mourad et al AIDS 2015[2] Wensing et al. Topics in antiretroviral Medicine 2016[3] Villabona et al. AIDS 2016[4] Palczewska et al. IEEE 2013