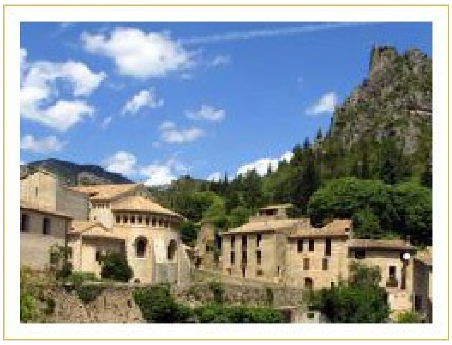


Mathematical and Computational Evolutionary Biology

May 27-31, 2013

GdR BiM

cnrs

UM1 Université Montpellier 1

um2 UNIVERSITÉ MONTPELLIER 2 SCIENCES ET TECHNIQUES

la Région Languedoc Roussillon

IBC Institut de biologie computationnelle

Labex NUMEV

Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier LIRMM

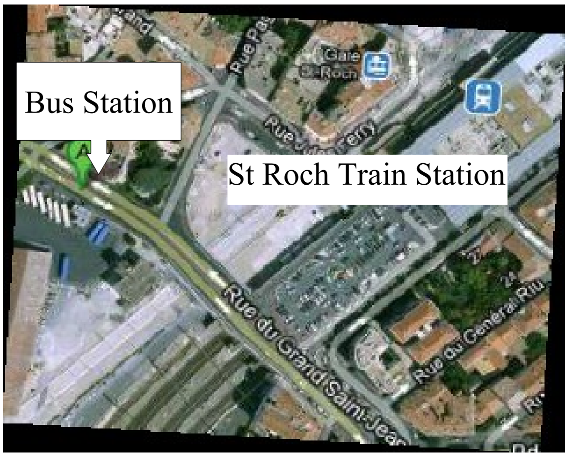
isem Institut des Sciences de l'Evolution-Montpellier

MiVEGEC Maladies Infectieuses et Vecteurs Ecologie Génétique Evolution et Contrôle

MATHématiqueS
en INteractions
pour la TERRE

INFORMATIONS

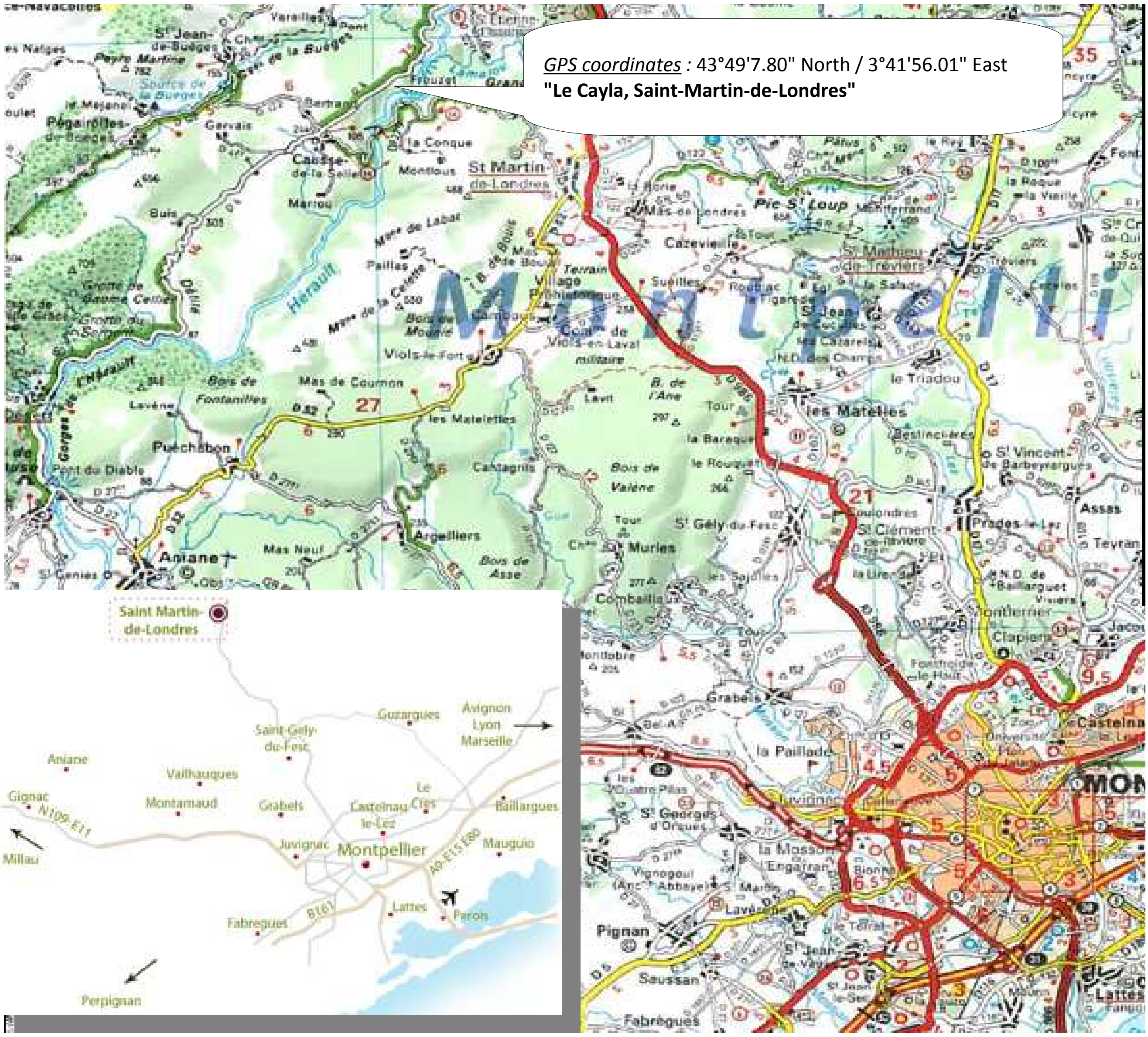
Meeting Point



Bus station, Parking du grand Saint-Jean
(Close to the train station and served by the tram)
A « Bancarel » bus will leave Montpellier on Monday at 9H00.

Location

The conference will be held at the ***Hameau de l'Etoile***,
a hamlet dedicated to seminars and conferences,
located at about 25 km north of Montpellier (south of France).



<u>GPS coordinates</u> *:* 43°49'7.80" North / 3°41'56.01" East
"Le Cayla, Saint-Martin-de-Londres"

Practical informations

Domaine Le Hameau de l'Etoile
Route de Frouzet
34380 ST-MARTIN-DE-LONDRES
Tél (+33) **04 67 55 75 73**
Fax (+33) 04 67 55 09 10

Taxi :

Taxi de St Martin de Londres
⇨ 20% discount for customers of Hameau de l'Etoile
Call first « Bernard » at **06 81 16 93 75**
Rates : In week, tram station Saint-Roch Train Station = 55 € / Airport=70 €
Week - end / Night : + 15 euros

Hotels in Montpellier :

Hôtel d'Aragon *** 10, rue Baudin 34000 MONTPELLIER Tél : 33 (0)4 67 10 70 00 fax : 33 (0)4 67 10 70 01	Hôtel d'Angleterre ** 7, rue Maguelone 34000 MONTPELLIER Tél : 33 (0)4 67 58 59 50 fax : 33 (0)4 67 58 29 52	Hôtel le Mistral ** 25, rue Boussairolles 34000 MONTPELLIER Tél : 33 (0)4 67 58 45 25 / 33 (0)6 60 53 73 40 fax : 33 (0)4 67 58 23 95
Hôtel Le Guilhem *** 18, rue Jean Jacques Rousseau 34000 MONTPELLIER Tél : 33 (0)4 67 52 90 90 fax : 33 (0)4 67 60 67 67	Hôtel des Arceaux ** 33/35, boulevard des Arceaux 34000 MONTPELLIER Tél : 33 (0)4 67 92 03 03 fax : 33 (0)4 67 92 05 09	Hôtel Nova ** 8, rue Richelieu 34000 MONTPELLIER Tél : 33 (0)4 67 60 79 85 fax : 33 (0)4 67 60 89 06
Newhotel du Midi *** 22, boulevard Victor Hugo 34000 MONTPELLIER Tél : 33 (0)4 67 92 69 61 fax : 33 (0)4 67 92 73 63	Hôtel des Arts ** 6, boulevard Victor Hugo 34000 MONTPELLIER Tél : 33 (0)4 67 58 69 20 fax : 33 (0)4 67 58 85 82	Hôtel du Palais ** 3, rue du Palais 34000 MONTPELLIER Tél : 33 (0)4 67 60 47 38 fax : 33 (0)4 67 60 40 23
Royal Hotel *** 8, rue Maguelone 34000 MONTPELLIER Tél : 33 (0)4 67 92 13 36 fax : 33 (0)4 67 92 59 80	Hôtel Colisée Verdun ** 33, rue de Verdun 34000 MONTPELLIER Tél : 33 (0)4 67 58 42 63 fax : 33 (0)4 67 58 98 27	Hôtel du Parc ** 8, rue Achille Bégé 34000 MONTPELLIER Tél : 33 (0)4 67 41 16 49 fax : 33 (0)4 67 54 10 05
Hôtel Acapulco ** 445, rue Auguste Broussonnet 34090 MONTPELLIER Tél : 33 (0)4 67 54 12 21 fax : 33 (0)4 67 52 26 10	Hôtel François de Lapeyronie ** 80, rue des Pétètes 34090 MONTPELLIER Tél : 33 (0)4 67 52 52 20 fax : 33 (0)4 67 63 56 65	Hôtel Les Troenes ** 17, avenue Emile Bertin Sans 34040 MONTPELLIER Tél : 33 (0)4 67 04 07 76 / 33 (0)6 29 02 31 17 fax : 33 (0)4 67 61 04 43
Hôtel Les Alizés ** 14, rue Jules Ferry 34000 MONTPELLIER Tél : 33 (0)4 67 12 85 35 fax : 33 (0)4 67 12 85 30	Hôtel Littoral ** 3, Impasse Saint Sauveur 34000 MONTPELLIER Tél : 33 (0)4 67 92 28 10 fax : 33 (0)4 67 92 72 20	Hôtel les Fauvettes * 8, rue Bonnard 34000 MONTPELLIER Tél : 33 (0)4 67 63 17 60 / 33 (0)6 89 26 63 58

PROGRAM

Monday, May 27

> 09h00 : **Meeting point** - Bus from Montpellier to Hameau de l'Etoile
> 10h00 : **Accueil** - Café & Croissants

> 11h00 – 12h30 : *KEYNOTE* ***p.10***
Tanja Stadler (ETH Zürich, CH)
« Phylogenetics in action - merging epidemiology and evolutionary biology »

> 12h30 : Déjeuner – Installation

> 14h15 – 16h30 : 6x 20 min *TALKS* (including questions)

Michael Blum (Laboratoire TIMC-IMAG, France) p.12
« Bayesian robust principal component analysis to detect genomic regions involved in local adaptation »

Miroslav Bacak (Max Planck Institute, Germany) p.11
« Computing medians and means of phylogenetic trees »

Trevor Bedford (University of Edinburgh, UK) p.12
« Antigenic flux in the influenza virus population »

Pause (15 min)

Michael Hiller (Max Planck Institute, Germany) p.13
« Linking phenotypic differences between species to differences in their genomes »

Alain Guénoche (IML, Université de Marseille, France) p.13
« Methods for consensus of partial trees »

Elina Numminen (University of Helsinki,Finland) p.16
« How much do the genotypes of bacteria reveal about the transmission events, when the majority of infections are unobserved? Assessing the probability of direct transmission with a branching process approach »

> 16h30 : Thé ou Café & gâteaux

> 17h00 – 18h30 : *KEYNOTE* ***p.8***
Steven Kelk (Maastricht University, NL)
« Recent advances in rooted phylogenetic networks: the long road to explicit hypothesis generation »

> 20h00 : Apéritif
> 20h30 : Dîner

Tuesday, May 28

> 09h00 – 10h30 : *KEYNOTE* ***p.7***
Niko Beerenwinkel (ETH Zürich, CH)
« Cancer as an evolutionary process »

> 10h30 : Thé ou Café

> 11h00 – 12h30 : *KEYNOTE* ***p.7***
Bastien Boussau (University of California, Berkeley, US)
« Genome-scale phylogenomics »

> 12h30 : Déjeuner

> 14h15 – 16h30 : 6x 20 min *TALKS* (including questions)

Maria Anisimova (ETH Zürich, CH) p.11
« Advantages of using Markov codon models for phylogeny reconstruction »

Roland F Schwarz (European Bioinformatics Institute, Cambridge) p.17
« MEDICC : Minimum Event Distance for Intra-tumour Copy number Comparisons »

Stephan Peischl (Institute of Ecology and Evolution University of Bern CH) p.17
« The accumulation of deleterious mutations during range expansions »

Pause (15 min)

Amaury Lambert (UPMC Univ Paris 06 and Collège de France, Paris) p.14
« Epidemics with random sampling »

Sarah Parks (EBI, Hinxton, Cambridge) p.16
« Non-Reversible Models for Phylogenetics Using Both Nucleotide and Amino Acid Data »

George Shirreff (ETH Zürich, CH) p.18
« The phyloanatomy of early SIV infection »

> 16h30 : Thé ou Café & gateaux - Balade, piscine....

> 19h00 – 20h30 : *POSTERS (even numbers)* ***p.19***
Vin, apéritif et discussions

> 20h30 : Dîner

Wenesday, May 29

> 09h00 – 10h30 : *KEYNOTE* ***p.9***
Gil Mc Vean (University of Oxford, UK)
« Dissecting the genetic contribution to human disease »

> 10h30 : Thé ou Café

> 11h00 – 12h30 : *KEYNOTE* ***p.8***
Ian Holmes (University of California, Berkeley, US)
« Phylogenetics grammars and heterogeneous space-time models »

> 12h30 : Déjeuner
> 14h00 - 20h00: Visite de Saint-Guilhem-le-Désert (Balade, Canoë ...)
> 20h30 : Dîner

Thursday, May 30

> 09h00 – 10h30 : *KEYNOTE* ***p.8***
Alexei Drummond (University of Auckland, NZ)
« Bayesian molecular epidemiology »

> 10h30 : Thé ou Café

> 11h00 – 12h30 : *KEYNOTE* ***p.9***
Darren Martin (University of Cape Town, SA)
« Factors influencing recombination in viruses »

> 12h30 : Déjeuner

> 14h15 – 16h30 : 6x 20 min *TALKS* (including questions)

Sivan Leviyang (Georgetown University, Washington DC) p.15
« Computational Inference Methods for HIV Escape from Immune System Response »

Sophie Lebre (Université de Strasbourg, Icube, France) p.14
« A new molecular evolution model for Limited Insertion Independent of Substitution (LIIS) »

Matthew Hartfield (Laboratoire MIVEGEC, France) p.13
« Determining the effect of Hepatitis C genotype on infection outcome »

Pause (15 min)

Peter Arndt (Max Planck Institute for Molecular Genetics, Berlin, Germany) p.11
« Neutral evolution of duplicated DNA - an evolutionary stick-breaking process causes scale-invariant behavior »

Filip Bielejec (Rega Institute for Medical Research, Belgium) p.12
« Inferring large-scale heterogeneous evolutionary processes through time »

Samantha Lycett (Ashworth Laboratories, University of Edinburgh, UK) p.15
« Bayesian methods for detecting epistatic interactions and compensatory mutations in Influenza A viruses »

> 16h30 : Thé ou Café & gâteaux - Balade, piscine....

> 19h00 – 20h30 : ***POSTERS (odd numbers)*** ***p.19***
Vin, apéritif et discussions

> 20h30 : Dîner

Friday, May 31

> 09h00 – 10h30 : ***KEYNOTE*** ***p.7***
Sebastian Bonhoeffer (ETH Zürich, CH)
« The paradox of heritability of HIV viral load? »

> 10h30 : Thé ou Café

> 11h00 – 12h30 : ***KEYNOTE*** ***p.9***
Erick Matsen (Fred Hutchinson Cancer Research Center, Seattle, US)
« Phylogenetics and the human microbiome »

> 12h30 : Déjeuner
> 14h30 : Bus to Montpellier

KEYNOTE SPEAKERS

> Niko Beerenwinkel
ETH Zürich, CH

Cancer as an evolutionary process

Cancer is a somatic evolutionary process driven by mutation and selection in an asexually reproducing population of tumor cells. We review recent developments in mathematical modeling of genetic cancer progression, including population genetics models, phylogenetic methods, and probabilistic graphical models of tumor progression. We discuss how cancer genome data obtained from massively parallel sequencing experiments can inform these models to gain insight into the evolutionary history of individual tumors and the evolutionary dynamics of cancer.

> Sebastian Bonhoeffer
ETH Zürich, CH

The paradox of heritability of HIV viral load?

Several studies have recently provided evidence that set point HIV load in donor and recipients are correlated arguing that virus load is at least partially under the genetic control of the virus (1-4). The observed high heritability of viral load, however, is difficult to reconcile with the following two relatively uncontroversial observations. First, set point virus load is relatively stable within a patient over prolonged periods during asymptomatic infection, but varies over several orders of magnitude cross-sectionally between patients (5). Second, viral evolution within an infected patient is rapid as is evidenced for example by the fast rates of immune escape, the rapid evolution of resistance and the rapid accumulation genetic diversity over the course of an infection (6). Thus, if we accept that virus load is heritable then the question arises as to how virus load can be stable over prolonged periods of time in a system that is known to have the capacity of rapid evolution. In other words, is it conceivable that virus load is a trait which is in part under the genetic control of the virus but that those factors that influence virus load are not linked to intra-host competitive ability? I have previously argued that differences in virus load may arise through differences in target cell activation rates (5). If target cell activation rate may be under the partial control of the virus, the factors responsible for target cell activation would be expected to evolve neutrally within the host. We present models of within and between host evolution in order to explore to what extent a hypothetical viral factor controlling target cell activation is compatible with the observed within patient stability, large cross-sectional variation and high heritability of virus load.

(1) S. Alizon et al, PLoS Path. 6, e1001123 (2010).
(2) F. M. Hecht et al., AIDS 24, 941–945 (2010).
(3) T. D. Hollingsworth et al., PLoS Path. 6, e1000876 (2010).
(4) J. Tang et al., AIDS Res. and Hum. Retrovir., 20, 19–25 (2004).
(5) S. Bonhoeffer et al., Trends Microbiol., 11, 499-504 (2003).
(6) A. Rambaut et al., Nature Rev. Genet. 5, 52-61 (2004).

> Bastien Boussau
University of California, Berkeley, US

Genome-scale phylogenomics

The histories of genes and species are tightly linked. Given a model of the relationship between gene and species phylogenies, knowledge of the species tree helps reconstructing gene trees, and knowledge of the gene trees helps reconstructing species trees. Unfortunately neither gene trees nor species trees can be observed in nature, and both need to be inferred, most often from sequence data. Assuming gene trees or gene tree distributions are given, several approaches have been developed to reconstruct species trees. Similarly, assuming species trees are given, several approaches have been developed to reconstruct gene trees. Finally, a few other approaches attempt to reconstruct jointly gene trees and species trees.

In this presentation I will first introduce the various models that have been used to describe the relationship between gene trees and species trees. I will focus more on models of gene duplication, transfer and loss than on coalescent models. I will then use results from the literature in cases where the species tree is given to show that these models improve gene tree reconstruction. Next I will present two maximum likelihood approaches for genome-size data sets: one for the coestimation of gene trees and species trees with a model of gene duplication and loss, and one for the estimation of species trees given gene trees with a model of duplication, transfer and loss. Finally I will introduce new work towards the Bayesian coestimation of gene trees and species trees using a model of duplication, transfer and loss.

> Alexei Drummond
University of Auckland, NZ

Bayesian molecular epidemiology

I will introduce the BEAST 2 software framework for Bayesian evolutionary analysis, focusing on its application to inference of epidemiological dynamics from molecular sequence data.

I will compare coalescent and birth-death-sampling based approaches to explaining epidemic behaviour, using some examples.

Then I will describe the machinery necessary to develop models for structured epidemics focusing on a recently implemented multi-type tree sampler that permits structured epidemic inference under either the structured coalescent or newly described birth-death-migration models. I will also introduce a new stochastic simulator for phylodynamic models that can provide prior expectations for all of the discussed models.

Finally I will describe direct-ancestor trees and discuss their use in modeling well-sampled epidemics. This leads naturally to a consideration of the distinction between epidemic transmission trees and viral gene trees and a general discussion on the outlook for research in Bayesian molecular epidemiology.

Overall I hope to provide an account of recent “phylodynamics research conducted by the Computational Evolution Group, and our close collaborators.

> Ian Holmes
University of California, Berkeley, US

Phylogenetics grammars and heterogeneous space-time models

Transformational grammars, as introduced formally by Noam Chomsky in 1956 (or Robert Lowth in 1762, if one goes further back) have proved to be a powerful tool in computer science, offering a mathematical formalism for describing and parsing the syntactic arrangement of natural language features (verbs, nouns, adjectives, etc). The stochastic versions of these grammars have proven quite useful in bioinformatics too, following the work in the 1990's on Stochastic Context-Free Grammars for RNA secondary structure analysis and Hidden Markov Models (aka Stochastic Regular Grammars) (Sean Eddy, Richard Durbin, David Haussler, et al).

As well as providing a "syntactic" model for the spatial organization of patterns in sequences, grammars can similarly be used to model the spatial organization of *aligned* sequences, in particular, sequences that have evolved on a phylogenetic tree under some common constraint. These sorts of "phylo-grammars" explain the spatial and temporal aspects of a sequence's evolutionary history, and have been used by genomic bioinformaticians to pick out all sorts of interesting signatures of selection, such as RNA structures overlapping protein-coding regions in viral genomes (Jotun Hein, Bjarne Knudsen, Jakob Skou Pedersen, Irmtraud Meyer et al) and "ultra-conserved" or "accelerated" genomic elements (David Haussler, Adam Siepel, Katherine Pollard et al).

My lab's software XRate http://biowiki.org/XRATE is something like an "interpreter" for phylo-grammars. That is to say, the user can specify the general hierarchically structured arrangement of "grammatical" features in an alignment, along with the general parametric form of the substitution models in the individual regions. The software then uses partially-supervised or unsupervised machine learning algorithms (specifically the Expectation Maximization algorithm) to estimate a maximum likelihood parametric fit to the data, and/or to annotate syntactic features. The numbers obtained can be used for downstream evolutionary analyses.

In this talk I will review the general phylo-grammar theory and present examples of several evolutionary analyses that have been performed with xrate.

> Steven Kelk
Maastricht University, NL

Recent advances in rooted phylogenetic networks: the long road to explicit hypothesis generation

In the last ten years interest has grown significantly in phylogenetic networks. As many researchers active in this field will testify, the name "phylogenetic network" actually reflects a wide array of sometimes quite different phylogenetic models. What unifies the models, however, is the idea that it is sometimes neither possible nor desirable to seek a single tree hypothesis to explain observed biological data. The field is also unified by the recognition that a scala of evolutionary phenomena can potentially distort (or render meaningless) the tree-building process. This includes genuinely "reticulate" phenomena such as horizontal gene transfer and hybridization (and, at the level of population genomics, recombination), but also phenomena such as incomplete lineage sorting and gene loss/duplication that are not inherently in contradiction with the tree model.

Looking beyond this common background the primary dichotomy within the field of phylogenetic networks is between what David Morrison has called "data-display phylogenetic networks" and "evolutionary phylogenetic networks". As their name suggests data-display networks do not attempt to generate an explicit hypothesis of history: they are an attempt to capture and summarise in a single picture all the places in the data where a tree model is not adequate to explain the patterns observed. Evolutionary phylogenetic networks, on the other hand, are much more ambitious: they generate a hypothesis of what happened, where, and how, a well-known example being the attempt to "superimpose" horizontal gene transfer events onto a tree-like backbone history.

The era of mainstream adoption of phylogenetic network models is not yet upon us. However, to the extent that they are used at all, data-display networks are far more popular than their evolutionary counterparts. The main reason for this is that data-display networks only summarise data, they do not (explicitly) impose a hypothesis on the user. Practicing biologists will only accept such an imposition if it resonates with their own (sometimes implicit) hypotheses and if the method has been emprically verified.

To become credible tools, therefore, evolutionary phylogenetic networks must first demonstrate their worth. They must demonstrate that they can be used to automatically generate evolutionary hypotheses that biologists themselves would eventually reach via more traditional, time-consuming, "manual" analysis.

In this talk I will give an overview of how far different evolutionary phylogenetic network models are progressing in terms of empirical verification. In how far can hypothesized horizontal gene transfer events be corroborated with acknowledged, "true" transfer events? In how far can hybridization models be used to predict the emergence of pathogenic hybrids? These are the type of questions I will look into.

> Darren Martin
University of Cape Town, SA

Factors influencing recombination in viruses

When two sufficiently related viruses replicate within the same cell it is possible that genomes will arise that contain genetic material from both viruses. The frequencies with which such recombinants arise, the genomic sites where recombination events occur and the fitness effects of these events can vary tremendously even between viruses in the same family. Whereas ecological factors will determine how frequently cells become co-infected with any particular pair of virus species/strains, biochemical factors will determine how frequently and at which genomic sites recombination events will occur. Following their generation, natural selection and chance will determine whether the descendants of recombinant genomes become common enough to be detected in viral genome sequence surveys. I will explain how to use virus genome sequence data both to compare recombination patterns of related viruses, and to identify the ecological, biochemical and selective factors contributing to these patterns.

> Erick Matsen
Fred Hutchinson Cancer Research Center, Seattle, US

Phylogenetics and the human microbiome

The human microbiome is the collection of microbes that live inside and on humans. The Human Microbiome Project (HMP) and the Metagenomics of the Human Intestinal Tract (MetaHIT) consortia have advanced human microboime research by generating thousands of 16s surveys and terabases of metagenomic data on the human microbiome, as well as funding bioinformatics and statistical development. In this talk I will review the impact of phylogenetics and tree-thinking on the methods used in the analysis of this data, as well as describing current challenges for phylogenetics coming from this type of work.

> Gil Mc Vean
University of Oxford, UK

Dissecting the genetic contribution to human disease

The opportunity to measure genetic variation across the whole genome of individuals with specific diseases has led, over the last ten years, to many insights into human disorders. These advances have been made possible by the joint development of new technologies that generate genome-scale data and statistical and computational tools for analysing and interpreting such information on a vast scale. In this lecture I will describe what we have learned about the structure of human genetic variation from large-scale studies including the HapMap and 1000 Genomes Projects, specifically focusing on the role of coalescent modelling in the analysis of recombination. I will also discuss the roles of natural selection and population history in influencing genetic variation in populations and describe some of the difficulties these cause for the detection of disease-associated variants. I will then describe some of the different experimental designs that are used to identify variants influencing disease risk, and statistical methods used to fine-map variants and to determine causal relationships between biomarkers and disease. Finally, I shall discuss some of the open problems in the analysis of genetic risk for disease.

> Tanja Stadler
ETH Zürich, CH

Phylogenetics in action - merging epidemiology and evolutionary biology
In my talk I will discuss phylogenetic concepts for inferring epidemiological parameters based on sequence data. Such inference is possible when pathogen evolution and epidemiology happens on the same time scale, and thus epidemiological processes leave a fingerprint in the pattern of genetic structure of pathogen populations. In particular RNA viruses evolve at the timescale on which their epidemic is happening, and thus viral sequence data provides a rich data source for inferring epidemiological processes. An increasing number of sequence data is becoming available due to routine drug resistance testing, and sophisticated statistical inference tools based on jointly modelling the evolution and epidemiology of viruses in a phylogenetic framework are developed. Thus, phylogenetic analyses now greatly add to our knowledge of epidemiological dynamics, which was before the revolution in sequencing technology mainly based on hard-to-collect incidence data. Sequence data in particular contains information about who infected whom, thus provides information about structured populations beyond the classic incidence data.

TALKS

> Maria Anisimova
ETH Zurich, Computer Science, CAB H82.2, Universitatstr. 6, Zurich, 8092

Advantages of using Markov codon models for phylogeny reconstruction

Compared to nucleotide or amino acid Markov models of substitution, codon models should provide a more realistic representation of protein-coding sequences since they naturally incorporate the structure of the genetic code and the selection intensity at the protein level. Thus for protein-coding genes phylogenetic inference is expected to be more accurate under codon models. Using CodonPhyML, our recent implementation of codon models for fast maximum likelihood inference of phylogenies, here I discuss the advantages advantages the codon models are able to provide. Reconstructed trees under amino acid, DNA and codon models are scrutinized for a set of real mammalian genes, focussing on statistical properties of inferred trees, distribution of branch lengths and changes in statistical branch supports.

> Peter Arndt [1], Florian Massip [2]
[1] Max Planck Institute for Molecular Genetics, Berlin, Germany
[2] INRA, Jouy-en-Josas, France

Neutral evolution of duplicated DNA - an evolutionary stick-breaking process causes scale-invariant behavior

Recently, an enrichment of identical matching sequences has been found in many eukaryotic genomes. Interestingly, their length distribution exhibits a power law tail raising the question of what evolutionary mechanism or functional constraints would be able to shape this distribution. We introduce a simple and evolutionarily neutral model, which involves only point mutations and segmental duplications, and produces the same statistical features in the length distribution of matching sequences as observed for genomic data. Further, we extend a mathematical model for random stick breaking to analytically show that the exponent of the power law tail is -3 and universal as it does not depend on microscopic details of the model. Evolutionary consequences for eukaryotic genomes are discussed.

> Miroslav Bacak, Philipp Benner
Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, 04103 Leipzig

Computing medians and means of phylogenetic trees

The tree space constructed in [Billera, Holmes, and Vogtmann: Geometry of the space of phylogenetic trees, 2001] as a model for phylogenetic trees, turns out to have very pleasant geometrical properties, namely, it is a polyhedral complex of nonpositive curvature. This space has recently attracted considerable attention of mathematicians, who extended a number of powerful optimization techniques we know from classical linear spaces into this nonlinear setting. After the discovery of a polynomial time algorithm for computing distances due to [Owen, Provan: A fast algorithm for computing geodesic distances in tree space, 2011], the tree space became highly demanded by computational biologists.

Among the most natural operations one wants to do with a given set of phylogenetic trees is computing its median and mean. It turns out that the classical algorithms do not function in the tree space and coming up with a new approach is a challenging mathematical problem.

In my talk, I will present novel algorithms for computing medians and means of a given set of trees, and explain their rigorous mathematical backgrounds as well as use in practice. Various versions (both deterministic and randomized) of these algorithms will be discussed and compared on a real data set.

Finally, I will put these algorithms into the perspective of statistical methods in phylogenetics.

> **Trevor Bedford**
University of Edinburgh

Antigenic flux in the influenza virus population

Owing to rapid mutation, the evolution of the influenza virus occurs on a human timescale; rather than being forced to infer past evolutionary events, we can observe them in near real-time. While individuals develop long-lasting immunity to particular influenza strains after infection, antigenic mutations to the influenza virus genome result in proteins that are recognized to a lesser degree by the human immune system, leaving individuals susceptible to future infection. Mutations are only transiently advantageous; the virus population must keep evolving antigenically to stay ahead of developing human immunity. This talk focuses on the process of antigenic innovation and the spread of novel virus strains. In this case, we have serological data from the hemagglutination inhibition (HI) assay that compares the level of cross-reactivity between different strains of influenza, as well as sequence data across strains. Here, we use a probabilistic framework called Bayesian multidimensional scaling (BMDS) to find a single consistent representation of antigenic distances between viruses by placing strains on a two-dimensional map. We integrate sequence evolution by treating BMDS location as a continuous diffusion across the phylogenetic tree. In this context, we examine the process of antigenic drift and investigate historical choices in vaccine strain by the World Health Organization.

> **Filip Bielejec**
Rega Institute for Medical Research, Minderbroedersstraat 10, BE-3000 Leuven

Inferring large-scale heterogeneous evolutionary processes through time

Molecular phylogenetic and phylogeographic reconstructions generally assume time-homogeneous substitution processes. Motivated by computational convenience, this assumption sacrifices biological realism and offers little opportunity to characterize the temporal dynamics of measurably evolving populations. We present a generic stochastic evolutionary model that relaxes the homogeneous substitution rate assumption by allowing the specification of different infinitesimal substitution rate matrices across different time intervals (called epochs) along the evolutionary history. This epoch model enables drawing inference about any discrete date type modeled as a continous-time Markov chain, including phylogeographic traits, where typically not only is the state-space large but the underlying rate matrix may also be non-reversible. We focus on an epoch model implementation in a Bayesian inference framework that offers great model flexibility. To alleviate the computational burden that the additional temporal heterogeneity imposes, we adopt a massively parallel approach that achieves fine-grain parallelization of the computations across branches that accommodate epoch transitions, making extensive use of graphics processing units. Through synthetic examples, we assess the models̉ performance in recovering evolutionary parameters from data generated according to various evolutionary scenarios that comprise different numbers of epochs for both nucleotide and codon substitution processes. We illustrate the usefulness of our inference framework in two different applications to empirical data sets: the selection dynamics on within-host HIV populations throughout infection and the seasonality of global influenza migration. In both cases, our epoch model captures key features of temporal heterogeneity that remained difficult to test using ad hoc procedures.
The methods used are available as part of the BEAGLE library for high performance statistical phylogenetics, along with suitable bindings to BEAST, a powerful user-friendly software package for performing Bayesian inference with molecular sequences via Markov chain Monte Carlo.

> **Michael Blum** [1], Nicolas Duforet-Frebourg [1]
[1] Laboratoire TIMC-IMAG, Domaine de la Merci, Faculté de Médecine, 38706 La Tronche France

Bayesian robust principal component analysis to detect genomic regions involved in local adaptation

Using large numbers of genomic markers, genome scans can reveal a proportion of loci that deviate from neutral expectations because they contribute to local adaptation (Foll and Gaggiotti 2008). This prominent biological process results in greater fitness of individuals in their local habitats due to natural selection. Understanding the genomic architecture of adaptation in humans is crucial to understand how past selection impacted disease susceptibility in modern populations (Barreiro and Quintana-Murci 2009). Here, we introduce an original method that seeks for outlier regions using Bayesian robust principal component analysis (BRPCA). Compared to more traditional approaches that are based on indices of genetic differentiation between populations, our approach is fully unsupervised and do not require populations to be defined in advance. With BRPCA, the algorithm learns the spatial directions under which local adaptation took place, which is desirable to infer the evolutionary history of adaptation (Coop et al. 2009). Using a large-scale human dataset, with hundreds of thousands of SNPs, we show the potential of the method and how it can scale to large genetic data set.

> Alain Guénoche [1]**,** Vladimir Makarenkov [2]
[1] IML, Université de Marseille
[2] UQAM, Montréal (Ca)

Methods for consensus of partial trees

X-Tree consensus is to summarize a set of X-trees in a single tree structure on X. It's a combinatorial problem admitting several solutions. To apply the Majority Rule, giving a median tree, all the X-trees must connect the same X set of leaves (or taxa). It is not always the case for gene trees. In this lecture we will present some algorithms to build a consensus tree from partial trees only connecting a subset of X. Some application to build a species tree from gene trees will be presented.

> Matthew Hartfield [1], Rowena Bull [2], Peter A. White [2], Andrew Lloyd [2], Fabio Luciani [2] and Samuel Alizon [1]
[1] Laboratoire MIVEGEC (UMR CNRS 5290, UR IRD 224, UM1, UM2) 911 avenue Agropolis, B.P. 64501, 34394 Montpellier Cedex 5, France;
[2] Evolutionary Dynamics of Infectious Diseases, School of Medical Sciences, University of New South Wales, Sydney, Australia

Determining the effect of Hepatitis C genotype on infection outcome

Infection by hepatitis C virus (HCV) leads to one of two outcomes, which varies amongst patients. Either the infection resolves itself within a matter of months, or it can persist over several years. It is difficult to ascertain to what extent this outcome is determined by the virus genotype itself using transmission networks, as these tend to be poorly known. Recently, phylogenetic methods have been created to estimate the proportion of set-point viral load that is inherited from one HIV-infected patient to the next. Studies found that up to half the variance in this trait is determined by the virus genotype. Here, we aim to investigate whether we can detect a similar signal in HCV infections. We first simulate inheritance of a binary trait outcome along a given phylogenetic tree to predict how traits gather in groups, and explain how these simulations are used to ascertain the virus effect on the infection outcome. Finally, we apply our method to HCV cohort data from Australia to try and detect an effect of virus genotype on whether hosts will clear the virus rapidly or develop a chronic infection. We also investigate whether key host SNPs, which are known to affect HCV infection outcome, affect this measure of inheritance.

> Michael Hiller
Max Planck Institute for Molecular Cell Biology and Genetics
& Max Planck Institute for the Physics of Complex Systems, 01307 Dresden, Germany

Linking phenotypic differences between species to differences in their genomes

Evolution has led to an amazing diversity in phenotypes between species. Since DNA as the blueprint of life encodes the phenotypes of an organism, phenotypic differences between species must be due to differences in their DNA. While the genomic era brings an unprecedented wealth of genotypic information, we know little about the genomic differences that explain the phenotypic diversity that we observe in nature. The reason is that any pair of species, even when as closely related as human and chimpanzee, exhibits millions of genomic changes and numerous phenotypic changes.

Here, we introduce a computational "forward genomics" strategy that - given only an independently lost phenotype and whole genomes - predicts phenotype-genotype associations for such traits. Our approach utilizes the fact that genomic regions encoding trait-specific information evolve neutrally and diverge faster in trait loss species. We quantify per-species sequence divergence in conserved genomic regions by reconstructing the likely DNA sequence of a common ancestor, taking great care to distinguish artifacts (such as assembly gaps and low-quality sequences) from real mutations. Forward genomics then matches genomic divergence and phenotypic loss patterns to associate specific regions with the given phenotype. We conducted genome-wide screens for two metabolic phenotypes. First, applied to "synthesis of vitamin C", a trait lost independently in primates, guinea pigs and many bats, forward genomics robustly pinpoints elevated sequence divergence in all vitamin C non-synthesizing lineages to the Gulo gene. Gulo encodes the final enzyme in vitamin C synthesis and we find it inactivated in all and only the known non-synthesizing lineages. Second, forward genomics attributed naturally low biliary phospholipid levels in guinea pigs and horses to the inactivated phospholipid transporter Abcb4. Human ABCB4 mutations also result in low phospholipid levels, but lead to severe liver disease, suggesting compensatory mechanisms in guinea pig and horse.

Genome-scale simulations show that forward genomics has substantial power and specificity to identify the genomic basis for these two traits and a large variety of other trait loss scenarios. Furthermore, we show that independently lost phenotypes are frequent. This suggests that forward genomics has broad applicability to other trait losses. Our approach will contribute to interpret the wealth of genome sequences to learn about evolutionary processes, natural diversity and also about the function of our genome and its contribution to human disease.

~ Hiller M et al. (2012): A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. Cell Reports, 2(4), 817-823

~ Hiller M, Schaar BT, and Bejerano G (2012): Hundreds of conserved non-coding genomic regions are independently lost in mammals. Nucleic Acids Res, 40(22), 11463 - 11476

> **Amaury Lambert** [1], P. Trapman [2]
A. Lambert [1], H. Alexander [3] and T. Stadler [3]
[1] UPMC Univ Paris 06 and Collège de France, Paris
[2] Stockholm University
[3] ETH Zürich

Epidemics with random sampling

We will present two works concerned with different characteristics of the same model of free growth epidemics where infected individuals are detected at some random time after their infection.
In the first work (coll. with P. Trapman), we characterize the state of the epidemics at the first time when some infective is detected. In the second work (coll. with H. Alexander and T. Stadler), we give the distribution of the phylogenetic tree spanned by detection points.

> **Sophie Lebre**, Christian Michel
Université de Strasbourg, Icube, 300 Boulevard Sébastien Brant 67400 Illkirch, France

A new molecular evolution model for Limited Insertion Independent of Substitution (LIIS)

We recently introduced a new molecular evolution model called the IDIS model for Insertion Deletion Independent of Substitution (Lèbre and Michel, 2010, 2012). In the IDIS model, the three independent processes of substitution, insertion and deletion of residues have constant rates. In order to control the genome expansion during evolution, we generalize here the IDIS model by introducing an insertion rate which decreases when the sequence grows and tends to 0 for a maximum sequence length nmax.

This new model, called LIIS for Limited Insertion Independent of Substitution, defines a matrix differential equation satisfied by a vector P(t) describing the sequence content in each residue at evolution time t. An analytical solution is obtained for any diagonalizable substitution matrix M. Thus, the LIIS model gives an expression of the sequence content vector P(t) in each residue under evolution time t as a function of the eigenvalues and the eigenvectors of matrix M, the residue insertion rate vector R, the total insertion rate r, the initial and maximum sequence lengths n0 and nmax, respectively, and the sequence content vector P(t0) at initial time t0. The derivation of the analytical solution is much more technical, compared to the IDIS model, as it involves Gauss hypergeometric functions.

Several propositions of the LIIS model are derived: proof that the IDIS model is a particular case of the LIIS model when the maximum sequence length nmax tends to infinity, fixed point, time scale, time step and time inversion. Using a relation between the sequence length l and the evolution time t, an expression of the LIIS model as a function of the sequence length l=n(t) is obtained. Formulas for `insertion only', i.e. when the substitution rates are all equal to 0, are derived at evolution time t and sequence length l. Analytical solutions of the LIIS model are explicitly derived, as a function of either evolution time t or sequence length l, for two classical substitution matrices: the 3-parameter symmetric substitution matrix (Kimura, 1981) (LIIS-SYM3) and the HKY asymmetric substitution matrix (Hasegawa et al., 1985) (LIIS-HKY).

An evaluation of the LIIS model (precisely LIIS-HKY) based on a statistical analysis of the GC content in complete genomes of four prokaryotic taxonomic groups, namely Chlamydiae, Crenarchaeota, Spirochaetes and Thermotogae, shows the expected improvement from the theory of the LIIS model compared to the IDIS model.

> Sivan Leviyang
Georgetown University, Department of Math. and Stat., St. Mary's Hall 3rd Floor, Washington DC, 20057

Computational Inference Methods for HIV Escape from Immune System Response

Early HIV infection is marked by immune response that targets multiple regions of the HIV genome. HIV escapes this selection through a large number of mutation pathways. In this talk I will describe a stochastic model of the complex selective sweep produced by such biology. The large number of mutation pathways makes the model high dimensional, requiring novel inference methods. I will describe computational inference methods that exploit HIV sequence data to explore the strength of immune response at different epitopes.

References :
Leviyang S., Computational Inference Methods for HIV-1 Selective Sweeps Shaped by Early Cytotoxic T-Lymphocyte Response. submitted. Preprint available on author website.
Leviyang S., The Coalescence of Intrahost HIV Lineages Under Symmetric CTL Attack. Bull. Math. Bio. v. 74, n.8, (2012), 509-535.
Leviyang S., Sampling HIV Intrahost Genealogies Based on a Model of Acute Stage CTL Response. Bull Math Biol. v 74, n.3, (2012) 509-35.

> Samantha Lycett, Mojca Zelnikar, Andrew Leigh Brown, Andrew Rambaut
Ashworth Laboratories, Institute of Evolutionary Biology, University of Edinburgh, Kings Buildings, Edinburgh, EH9 3JT, UK

Bayesian methods for detecting epistatic interactions and compensatory mutations in Influenza A viruses

The evolution of Influenza A viruses in humans is driven by the need to generate antigenically novel variants in order to escape population immunity. Furthermore a few strains can rapidly increase in frequency and dominate the viral population in particular seasons or years. In pandemic H1N1(pdm09) and seasonal H1N1 strains, several mutations are postulated to have potential virulence enhancing or other adaptive properties. However, mutations can be driven to moderate or high prevalence due to founder effects in a rapidly growing population rather than through associated increases in viral fitness. Consequently distinguishing neutral or mildly deleterious mutations from important fitness enhancing mutations, or combinations of mutations, in the host population is problematic.

Sequences sampled between 1996-2009 (seasonal H1N1) and 2009-2012 (pandemic H1N1) were used to infer sets of time resolved phylogenetic trees, together with viral effective population size as a function of time using BEAST (Bayesian Evolutionary Analysis by Sampling Trees). To investigate the detection of epistatic interactions and compensatory mutations, we compared two approaches: (1) Putative fitness enhancing amino acid mutations were represented as independent discrete traits, asymmetric rate models were fitted within BEAST, unobserved ancestral states were reconstructed upon a posterior sample of 1000 trees, and interactions were detected by calculating the association of changes from ancestral states for all tree branches for sites of interest. (2) Rate models in which combinations of mutations in pairs of sites of interest were also inferred, and compared to the independent models using AIC over the same set of trees.

Using a combination of simulations and data analysis we found that the first method could detect strong interactions, but was limited in power in our data sets since it relies on changes at both sites occurring on the same branch, and there being more than one occurrence of this effect. The second method enabled us to detect compensatory changes, where a mutation at one site was closely followed by a mutation at a second site. Using this method we found phylogenetic evidence for interactions within the Neuraminidase protein, and epistatic interactions between Neuraminidase and Hemagglutinin including the active site and antigenic sites. These results help explain how seemingly less fit strains containing drug resistant Neuraminidase (in the absence of drug use) can nevertheless rise to dominance in the population by association with Hemagglutinins with appropriate favourable mutations, and show how phylogenetic analyses may be used to give an estimate the probability of dangerous combinations of mutations occurring

> **Elina Numminen** [1], Jukka Sirén [1], Jukka Corander [1]
[1] Department of Mathematics and Statistics, University of Helsinki, P.O. Box 68, 00014 Helsinki, Finland

How much do the genotypes of bacteria reveal about the transmission events, when the majority of infections are unobserved? Assessing the probability of direct transmission with a branching process approach

Transmission trees that describe the epidemic spread from host to host, are of interest, since they might reveal about spatial spread, the utilities of intervention measures or differences between pathogenic strains, for instance. Such trees are rarely known, and they rather have to be inferred from data. Symptom onset times and genotypes of bacterial isolates are both informative on transmission tree. While Cottam et. al. (1.) used phylogenies of isolates to only rule out certain impossible transmission trees, genetic information could be directly incorporated in the likelihood of transmission tree, as shown by Ypma et al. (2.) and Jombart et al. (3.). Still, the two methods have difficulty in dealing with unknown sources of transmission. Either it is assumed that all the infected cases are observed, or then the method cannot tell whether the inferred transmission between a pair of hosts is actually direct. Assuming all individuals contributing to the epidemic to be observed is in many cases far-fetched. Also augmentation of the state-space to consider all the possible states of unobserved individuals can prove out to be very complex.

Thus in our current work, we consider a statistical inference framework for a situation where the majority of infected hosts were never observed and very well can be sources of infection of an observed case. Rather than constructing precise transmission trees, we study the probability of direct transmission between pairs of individuals we have observations about. We also assess the probability that the transmission came from an unobserved source in the local population. This is done by using a branching process model for modeling the stochastic features in the offspring distribution of an infection, where the properties of the branching process are adjusted to match the epidemiological characteristics of the transmission process observed in the data, or known a priori. This allows us to assess the posterior probability of direct and indirect transmission between two infected individuals, given the observations; the genetic distances and the times of observing the isolates.

We apply our methods for analysis of between-household transmission of an endemic bacteria in a refugee camp. Based on our preliminary results, we conclude that both temporal and genotypic information have tremendous influence on how probable the direct transmission is and the impact the two is not obvious from the raw data, let alone from our a priori guesses.

References:
1. Cottam E.M. et al. Intagrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus, Proc R Soc B, 275: 887-895
2. Ypma R. J. F. et al. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data, Proc R Soc B, 2012, 279: 444 - 450.
3. Jombart T. et al. Reconstructing disease outbreaks from genetic data: a graph approach, Heredity 2011, 106: 383-390.

> **Sarah Parks**, Nick Goldman
EBI, Hinxton, Cambridge

Non-Reversible Models for Phylogenetics Using Both Nucleotide and Amino Acid Data

Almost all evolutionary models commonly used in phylogenetic analysis are a subset of the general time reversible (GTR, REV) model. These models all assume the mathematical condition of time reversibility. This assumption was introduced to reduce computational effort and to ease mathematical complexity when calculating likelihoods; it has no biological basis. Relaxing this assumption by using a non-reversible model may fit data significantly better, potentially giving a more accurate description of evolution, better trees and other benefits.

A consensus has not been reached on whether non-reversible models are significantly better than reversible models. Often studies have only looked at a few trees and alignments and made decisions on the utility of non-reversible models based on this. Further, all of this work has been done on nucleotide datasets. No previous research has investigated non-reversible models for amino acid datasets, even though many phylogenies are built using amino acid data.

We have explored the use of non-reversible models for both amino acid and nucleotide datasets. This involves use of likelihood ratio tests between reversible and non-reversible models. Additionally we are interested in measures of non-reversibility, i.e. metrics of the deviation of a non-reversible model from reversibility. We have devised a number of different measures of reversibility and explored their relationship to the likelihood ratio test.

The results to be presented show that for nucleotide data non-reversible models are often a significantly better description of the evolutionary process than reversible models. The branch lengths of resulting inferred trees however are not significantly changed by the use of a non-reversible model. For amino acid data a much smaller proportion of the datasets studied deviate significantly from reversibility. This is probably due to more data being needed to determine significance for amino acids than for nucleotides.

The measures of non-reversibility show that non-reversible models range from being almost reversible to very non-reversible. This correlates with the results of the likelihood ratio test, but not strongly. As we expected, the measures are assessing the degree of non-reversibility whereas the likelihood ratio test assesses the strength of evidence of non-reversibility. While the degree of non-reversibility contributes to the evidence of non-reversibility, the amount of data also has the predicted effects and we show that these two things are distinct and measurable.

> **Stephan Peischl**, Isabelle Dupanloup, Mark Kirkpatrick, Laurent Excoffier
[1] Institute of Ecology and Evolution University of Bern Baltzerstrasse 6 CH-3012 Bern Switzerland
[2] Section of Integrative Biology The University of Texas at Austin 1 University Station #C0930 Austin, TX 78712

The accumulation of deleterious mutations during range expansions

Recent studies have shown that the ecology, the genetics, and the biology of populations can be affected by range expansions, but the exact underlying processes at work are still largely unknown. Given that most species have gone through ancient or recent range expansions, a better theoretical understanding of these phenomena is needed. We used a combination of simulation and analytical methods to study the evolution of fitness during one- or two-dimensional range expansions under an influx of beneficial and deleterious mutations throughout the genome. We show that strong genetic drift and inefficient selection on the front of range expansions can lead to an accumulation of deleterious mutations and therefore to the build-up of what we call an expansion load. This expansion load occurs under most studied scenarios and it is a major component of the total mutation load in newly colonized territories even thousands of generations after the expansion has stopped. We also find that expansion load can affect the dynamics of the expansion and that it can even temporarily stop the expansion. The phenomenon of expansion load is compatible with and explains a growing body of evidence suggesting that populations that have recently expanded, e.g., humans, show an excess of deleterious mutations.

> **Roland F Schwarz** [1,2], Anne Trinh [2], Botond Sipos [1], James D Brenton [2], Nick Goldman [1], Florian Markowetz [2]
[1] *European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD;*
[2] *University of Cambridge, CRUK Cambridge Research Institute, Robinson Way Cambridge, CB2 0RE, UK*

MEDICC : Minimum Event Distance for Intra-tumour Copy number Comparisons

Tumour heterogeneity, i.e. the genomic diversity of cancer cells within a single tumour, is thought to be the source of chemotherapy resistance. In many cancers, this heterogeneity is not limited to point mutations but includes large scale genomic rearrangements and endoreduplications that lead to aberrant copy number (CN) profiles. Reconstruction of the evolutionary tree of cancer within the patient allows us to quantify and understand the aetiology of tumour heterogeneity.

In some cancers, such as high-grade serous ovarian cancer (HGSOC), CN profiles predominate. However tree inference is hindered by unknown phasing of major and minor CNs, horizontal dependencies between adjacent genomic loci and the lack of curated CN profile databases to use as a reference for probabilistic inference.

To address these problems we have developed MEDICC (Minimum Event Distance for Intra-tumour Copy number Comparisons), an algorithm for phylogenetic reconstruction based on CN profiles. MEDICC uses finite-state transducers to encode a minimum evolution criterion that determines pairwise evolutionary distances between CN profiles. This minimum-event distance computes the smallest number of amplification and deletions of arbitrary length that are necessary to transform one genomic profile into another. Using this approach we are able to phase major and minor CN profiles to the parental alleles and infer trees and ancestral genomes, thereby minimizing the overall tree length. The distance measure is formulated such that the resulting matrix of pairwise distances has a direct mapping to a positive semi-definite kernel matrix. This allows us to perform principal component analysis in evolutionary space and use this embedding to numerically quantify tumour heterogeneity and other quantities of interest, such as the degree of clonal expansion, using spatial statistics.

After extensive simulation and validation we applied our method to a novel clinical study of 20 multiply-sampled HGSOC patients, reconstructed evolutionary trees and quantified heterogeneity and the degree of clonal expansion. Our results show that the degree of heterogeneity and clonal expansion has a strong negative association with patient survival times. Ancestral reconstructions allowed us to pinpoint driver events in HGSOC such as loss of chromosome 17q and we were able to accurately place relapse samples in the evolutionary trees, showing that resistant relapse is a clonal expansion of a minor subclone already embodied in presentation disease.

MEDICC is the first rigorous tree inference algorithm for CN profiles that makes full use of the bi-allelic frequencies obtained from SNP arrays. Its minimum event distance allows accurate reconstructions of phylogenies without the need for rearrangement probabilities. The definiteness properties of the computed distance matrices allow for the direct application of numerous machine learning algorithms, which include classification, exploration and regression, in this evolutionary space. Its scope is not limited to cancer genomics or SNP arrays but covers the general case of estimation of rearrangement phylogenies from CN data. In this sense MEDICC closes the gap between traditional methods, such as the works of Sankoff and Pevzner, which mostly start from identifiable individual genomic segments, and the bulk of rearrangement data available, which are aggregated CNs per segment without positional information.

> **George Shirreff** [1], Victor Garcia [1], Thomas H. Vanderford [2], Guido Silvestri [2], and Roland R. Regoes [1]
[1] Theoretical Biology, ETH, Universitätstrasse 16, 8006 Zürich, Switzerland
[2] Yerkes National Primate Research Center, Emory University, Atlanta, Georgia, USA

The phyloanatomy of early SIV infection.

Introduction :

The early stages of infection with HIV are very difficult to study due to the virus' low detectability and uncertainty as to the moment of infection. However, understanding this early stage is also very important for understanding key phenomena which occur early in infection such as the transmission bottleneck, the rapid depletion of gut associated lymphocytes, the spread into the latent reservoir, the rise of the early immune response, and the corresponding viral escape response. A crucial issue which is relevant to all of these is the extent to which virus migrates between different types of tissue and what their different degrees of diversity, in general and with respect to specific mutations.

Methods :

We address these questions by studying data collected by next generation (454) sequencing from 15 rhesus macaques experimentally infected with SIV. Samples were taken at multiple time points (7 to 168 days) and from multiple tissue compartments (lymph nodes, rectal biopsy, peripheral blood mononuclear cells and plasma). We examine the change in diversity between compartments and time points, and where particular mutations arise. We estimate the rates of migration between different body compartments, using techniques derived from phylogeography.

Results :

We generally observe an increase in diversity over time, but a drop in diversity is often observed later in infection which corresponds to the sweep of escape mutations through the population. In agreement with previous analysis of this dataset, escape mutations are most often observed to occur first in the lymph nodes. The migration rates between different body compartments are highly variable between animals, which may reflect host-based differences in the course of infection, or may be the result of stochastic effects.

Conclusion :

Next generation sequencing data provide many opportunities to study phyloanatomy in early infection. Using these kind of data also present challenges. A modeling based approach to simulate the system would allow us to estimate these rates of migration and other phenomena in detail, while also incorporating uncertainty.

POSTERS

Poster 1

> **Samuel Alizon** [1], Olivier Gascuel [2,3] , Guilhem Heinrich [1], Matthieu Jung [3]
[1] MIVEGEC, IRD Montpellier, France
[2] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM), UMR5506 CNRS, Université Montpellier 2, France
[3] Institut de Biologie Computationnelle (IBC), 95 Rue de la Galéra, 34095 Montpellier, France

Linking epidemiological dynamics and phylogenies

There is a growing interest in using phylogenies inferred from pathogens sequences (especiallly viruses) to estimate epidemiological parameters. However, most existing methods are restricted to very simple epidemiological settings (e.g. SIR models, where the number of susceptible hosts is assumed to be constant). Our goal is to develop a general framework that allows us to use phylogenetic data not only to infer parameters for any epidemiological model, but also to compare epidemiological models. The first step in our task is to determine whether different epidemiological models lead to phylogenies with significantly different shapes. To this end, we constructed an individual-based simulation model that allows us to build a phylogeny for a stochastic run of any type of epidemiological model. We then compare the power of various summary statistics to detect parameter variations in the same model. Finally, we use our summary statistics to test our ability to compare different models.

Poster 2

> **Sarah Bastkowski** [1], Andreas Spillner [2], Vincent Moulton [1]
[1] School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK,
[2] Institute of Mathematics and Informatics, University of Greifswald, Walther-Rathenau-Straße 47, 17487 Greifswald, Germany

Fishing for Minimum Evolution Trees with Neighbor-Nets

Phylogenetic trees are used by biologists to represent the evolutionary history of a set of species. A common approach to construct phylogenetic trees is to search through the space of all such trees for one that optimizes some score function, such as the minimum evolution criterion. As this can be difficult in general, a possible alternative approach suggested by David Bryant is to instead search for trees within sets of bipartitions or splits of the set of species in question. Here we consider the problem of searching through a set of splits that is circular. Such sets can be generated by the neighbornet algorithm for constructing phylogenetic networks. We show that this can be done in O(n^4) for a set of species of size n, and through simulations compare our approach with FastME, a leading method for contructing minimum evolution trees which searches through tree space using tree operations. Our results indicate that even though a circular split system represents just a small fraction of all possible splits, if we construct them with the neighbornet algorithm, they seem to capture relevant information for the construction of minimum evoluton trees.

Poster 3

> **Philipp Benner**, Miroslav Bacak, Pierre-Yves Bourguignon
Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, 04103 Leipzig

Inference of Phylogenetic Trees

The inference of phylogenetic trees from multiple sequence alignments is key to many problems in computational biology. Applications, such as the analysis of ChIP-Seq data for motif discovery or the identification of conserved sites, rely on a sound estimate of the tree. From a decision theoretic perspective we should use as point estimate the phylogenetic tree that corresponds to the posterior expectation. Usually, the tree topology is subject to uncertainty, which requires that we have a clear definiton of a tree space on which the expectation is computed. Fortunately, such a space was recently introduced by Billera et al. [2001]. The posterior probability of a tree could not be evaluated without a model for multiple sequence alignments. For this, a phylogenetic tree is complemented with an evolutionary model. Much dispute exists on how alignments of non-functional sites should be modeled. Such regions show a very low level of conservation. Two main approaches can be found in the literature.

One approach, taken for instance by Siepel et al. [2005], uses a separate phylogenetic tree with scaled branch lengths for non-conserved sites. Another approach is to drop the phylogeneticdimension and to assume full independence between sequences (see for instance Siddharthan et al. [2005]). So far, no clear argument exists for either approach. We provide a full generative model for multiple sequence alignments that allows a clear interpretation of branch lengths in phylogenetic trees. It differs in essential details from models that are currently used to estimate trees (e.g. Huelsenbeck and Bollback [2001], Ronquist and Huelsenbeck [2003]). A thorough interpretation of the model shows that branch lengths should in fact be independent of the level of conservation by using a suitable site-dependent stationary distribution of the evolutionary model. Therefore, a single tree is sufficient to model both conserved and non- conserved sites. Our approach requires to integrate over the stationary distribution, for which we provide an exact algorithm. Inference methods heavily rely on the use of Markov chain Monte Carlo (MCMC) sampling. The posterior expectation can only be approximated by averaging over the posterior samples which is aggravated by the fact that samples might have varying tree topology. However, recent developments in Hadamard spaces and the introduction of a method to compute the Fréchet mean by Bačák [2012, 2013], Miller et al. [2012] allow us to obtain the posterior expectation. By combining those methods, we present a full picture of how phylogenetic trees can be inferred from multiple sequence alignments. Our findings suggest that estimates of branch lengths tend to be too short. Whether or not this also affects the estimated tree topologies is yet to be determined.

References
~ Miroslav Bacak. A novel algorithm for computing the fréchet mean in hadamard spaces. preprint arXiv:1210.2145, 2012, 2012.
~ Miroslav Bacak. Computing medians and means via the proximal point algorithm. Submitted, 2013.
~ Louis J. Billera, Susan P. Holmes, and Karen Vogtmann. Geometry of the Space of Phylogenetic Trees. Advances in Applied Mathematics, 27:733–767, 2001. doi: 10.1006/aama.2001. 0759.
~ John P. Huelsenbeck and Jonathan P. Bollback. Empirical and hierarchical bayesian estimation of ancestral states. Systems Biology, 50(3):351–366, 2001.
~ E. Miller, M. Owen, and S. Provan. arXiv:1211.7046v1, 2012. Averaging metric phylogenetic trees. Preprint,
~ Fredrik Ronquist and John P. Huelsenbeck. Mrbayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics Applications Note, 19(12):1572–1574, 2003.
~ Rahul Siddharthan, Eric D. Siggia, and Erik van Nimwegen. Phylogibbs: A gibbs sampling motif finder that incorporates phylogeny. PLoS Computational Biology, 1(7), 2005.
~ A. Siepel, G. Bejerano, Js, As, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, Lw, S. Richards, Gm, Rk, Ra, Wj, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research, 15(8):1034–1050, 2005.

Poster 4

> **Manuel Binet** [1,2], Olivier Gascuel [1,2], Céline Scornavacca [2,3], Emmanuel J.P. Douzery [3] and Fabio Pardi [1,2].
[1] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM), UMR5506 CNRS, Université Montpellier 2, France;
[2] Institut de Biologie Computationnelle (IBC), 95 Rue de la Galéra, 34095 Montpellier, France
[3] Institut des Sciences de l'Evolution de Montpellier (ISEM), UMR5554 CNRS, Université Montpellier 2, France

Distance-based phylogenomics

As an alternative to the main approaches that have been proposed in phylogenomics to infer phylogenies from multiple genes, we propose a new method, RBLE (Rates and Branch Lengths Estimation). This is a distance-based method whose goal is to reconstruct the tree that best fits a collection of distance matrices, each one estimated from a different gene present in a subset of the studied organisms. Least-squares criteria and linear algebra procedures are used to obtain low computing times. We have implemented RBLE and tested it on simulated data, obtaining encouraging results. Because of the linear complexity in the number of matrices in input, our method looks particularly suitable for the large datasets available today, where the number of genes is in the order of the thousands.

Poster 5

> **Yao-ban Chan** [1], Céline Scornavacca [1], Vincent Ranwez [2]
[1] ISEM, Université Montpellier 2
[2] Montpellier SupAgro

TL or not TL: Reconciliation spaces and similarities

Increasing interest has been focusing recently on the problem of gene tree/species tree reconciliations. Given a binary tree representing the evolutionary history of the species and a binary tree representing the evolution of a gene family of those species, reconciliation seeks to infer the ``hidden" genetic evolutionary events of duplication, transfer and loss that explain apparent discrepancies between those two trees. In this talk, we introduce a set of elementary reconciliation operators that allows us to explore the full space of all possible reconciliations between a gene and species tree. This allows us to analyse the nature of the space, derive an analytical formula to count those reconciliations, a practical way to explore the neighborhood of a given one (a pre-requisite for meta-heuristics such as hill-climbing or simulated annealing) and a formal definition of similarity and distance between two reconciliations.

Poster 6

> **Benny Chor**, Jonathan Witztum, Erez Persi, David Horn, and Metsada Pasmanik-Chor
School of Computer Science, Tel-Aviv University

Metazoan Conservation Profiles, Functional Enrichment, and the Tree of Life

The availability of many complete, annotated proteomes, as well as gene annotation and enrichment tools, enables the systematic study of the relationships between protein conservation and functionality.

We explore this question based solely on the presence or absence of protein homologues, namely the conservation profiles. We examine the proteomes of 18 metazoans, from two distinct points of view: the humans and the flys, and study relations between protein conservation profiles, functionality, and evolutionary history as represented by the tree of life.

Poster 7

> **Miraine Davila** [1], Amaury Lambert [1], Bernard Cazelles [2]
[1] UPMC, 4 place Jussieu, Tour 16-26 1er étage, 75005 PARIS;
[2] UMR 7625, UPMC-CNRS-ENS, 46 rue d'Ulm, 75230 Paris Cedex 05

Inferring population dynamics from virus phylogenies: likelihood computation

To understand the emerging new pathogens affecting different populations, it is now important to consider the connections between epidemiological and evolutionary mechanisms [2]. Mathematical modeling can provide a theoretical framework to improve our understanding of these complex and constantly evolving systems in order to achieve innovative approaches for their detection and characterization.

In recent years the amount of works involving the use of phylogenies of extant taxa to infer the patterns of diversification has considerably grown. After the first likelihood based-method to infer speciation and extinction rates on the basis of reconstructed birth-death process presented in [3] , many authors have developed similar ideas under different scenarios. In particular, the availability of viral sequences allows the increase of applications of such models in the field of epidemiology [4], [5].

When the population of susceptible individuals is sufficiently large, one can be placed in the context of a simplified model without density dependence, the previously mentioned birth-death process with constant rates. However, even under this (simple) model, the quantification of the likelihood on the basis of available data can be a delicate and complex issue. Especially when one considers there is lack of information at the present, or even more if we dispose of some extra information about the past, such as the total extant population at earlier moments.

The present work is focused in obtaining the likelihood of an incomplete sampled population (available epidemiological data) on present time, jointly with their reconstructed phylogeny, and with the population size in several (deterministic) past times, and conditional on the survival of the population, issued from a single individual at time 0. This distribution is characterized through its multidimensional probability generating function, obtained thanks to the contour process described in [6], [7] and a series of inhomogeneous branching processes observed when we regard the population backward in time. The likelihood computation can be then achieved numerically, allowing parameter estimation.

References:
[2] Wolfe ND, Dunavan CP, Diamond J (2007). Origins of major human infectious diseases. Nature 447: 279-283
[3] Nee S, May RM, Harvey PH, 1994. The Reconstructed Evolutionary Process. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 344(1309):305-311.
[4] Volz EM, Pond SLK, Ward MJ, Brown AJL, Frost SDW 2009. Phylodynamics of infectious disease epidemics. Genetics 183: 1421-1430.
[5] Lambert A, Alexander H, Stadler T, 2012. Sampling epidemics with general lifetimes through time (submitted).
[6] Popovic L, 2004. Asymptotic genealogy of a critical branching process. Ann. Appl. Probab., 14(4):2120-2148.
[7] Lambert A, 2010. The contour of splitting trees is a Lévy process. Ann. Probab., 38(1):348-395.

Poster 8

> **Nicola De Maio**, Carolin Kosiol
VetMedUni Vienna, VeterinarPlatz 1, Wien, Austria

Polymorphism-aware Phylogenetic Models

Comparative analysis of related genomes are commonly performed to detect sites affected by selection, that can slow or accelerate evolution. Genome-wide scans have reported that genes involved in the immune system are enriched in accelerated evolution, consistently with the red queen hypothesis. However, it has also been shown that spurious signals of positive selection might come from selection on synonymous sites, as well as from neutral processes such as biased gene conversion. It is therefore important to distinguish between variation in neutral evolutionary rates (mutation and biased gene conversion primarily) and selection.

Standard phylogenetic models represent substitutions (new alleles introduced via mutations and then fixed in the population), as instantaneous events. Because of this approximation, it is usually hard to disentangle mutation rates and fixation biases in comparative analysis. Furthermore, using simulations, we show that standard models suffer from biases in estimation of short phylogenetic branches due to incomplete lineage sorting and ancestral polymorphisms.

We propose a new POlymorphisms-aware phylogenetic MOdel (PoMo) that relaxes the assumption of instantaneous substitutions. PoMo is a phylogenetic Markov model with states representing fixed alleles (as standard nucleotide models) as well as states representing polymorphisms at different allele frequencies. A substitution is hereby obtained through a mutational event followed by a gradual fixation process. Polymorphisms can either be observed in the present (tips of the phylogeny) or be ancestral (present at inner nodes). Our model utilizes both divergence and polymorphism data from different species/populations.

We analyze genome-wide synonymous sites of human, chimpanzee, and two orangutan species alignments. For each taxon we include data from several individuals. Using PoMo, we obtain accurate estimates of mutation rates, and of the intensity of biased gene conversion (BGC) that acted on Great Apes since their split. Furthermore, PoMo is non-stationary, and therefore we can estimate equilibrium and root nucleotide frequencies. Equilibrium frequencies strongly depend on present GC content, but overall GC content is homogenizing. We confirm that mutation rates are significantly context-dependent and strand-specific. We also find that both mutation rates and BGC vary with GC content, determining differences in substitution rates among isochores.

In addition to investigating variation along the genome, we plan to study clade-specific evolutionary patterns, and in particular to estimate changes in mutation rates and BGC between the orang clade and the human-chimp clade. I will also show how our models can be used to test individual genes for signatures of selection.

Poster 9

> **Frédéric Delsuc** [1], Monsef Benkirane [2] and Nicolas Lartillot [3]
[1] Institut des Sciences de l'Evolution, UMR5554-CNRS-Université Montpellier 2, Montpellier, France
[2] Institut de Génétique Humaine, UPR1142-CNRS, Montpellier, France
[3] Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Québec, Canada

Molecular evolution of HIV restriction factors during Primates evolution using a new Bayesian method for estimating branch-specific dN/dS ratios.

The identification of HIV restriction factors is an active research area. Comparative evolutionary analyzes can provide valuable assistance to the characterization of these factors and their molecular mode of action. Indeed, their molecular evolution is characterized by a coevolution with viral antagonists. This evolutionary arm race between host and virus genes leaves detectable signatures in their sequences left by recurrent episodes of positive selection. The recently identified SAMHD1 is only the fifth HIV restriction factor to be discovered. It has been shown that the evolution of this gene was dictated by its interaction with the viral antagonist vpx gene found in the human HIV2 virus and in some primate SIVs (Laguette et al. 2012; Lim et al. 2012). Our detailed analyses of selection pressures have allowed determining the probable age of onset of the first selective pressures corresponding to the origin of infection by primate lentiviruses. In this work, we extended these analyses to two other HIV restriction factors (BST2 and TRIM5a) by applying a new Bayesian method for the characterization of selection pressures along the branches of a phylogenetic tree. The results of these comparative analyses that HIV restriction factors have experienced widely different history of selection pressures during primate evolution, probably reflecting their different modes of action.

References:
~ Laguette N, Rahm N, Sobhian B, Chable-Bessia C, Münch J, Snoeck J, Sauter D, Switzer WM, Heneine W, Kirchhoff F, Delsuc F, Telenti A, Benkirane M (2012). Evolutionary and functional analyses of the interaction between the myeloid restriction factor SAMHD1 and the lentiviral Vpx protein. Cell Host Microbe 11:205-17.
~ Lim ES, Fregoso OI, McCoy CO, Matsen FA, Malik HS, Emerman M (2012). The ability of primate lentiviruses to degrade the monocyte restriction factor SAMHD1 preceded the birth of the viral accessory protein Vpx. Cell Host Microbe 11:194-204.

Poster 10

> **Linda Dib**, Nicolas Salamin
Unil / Sorge Biophore 1015 Lausanne Switzerland

Modeling the evolution of co-evolving positions in the melanocortin system

Coevolution is defined as the change of a biological object triggered by the change of a related object, and it has been observed in the DNA and amino acid sequences. Ten years ago it has been demonstrated that evolutionarily co-evolving networks of residues in a protein mediate allosteric communication involved in cellular signaling. However, a newly published methodology named BIS show that co-evolving positions can also explain folding intermediates, peptide assembly, key mutations with known roles in genetic diseases, distinguished subfamily-dependent motifs and differentiated evolutionary pressures on protein regions.

Although predictions and biological evidence showed that some positions are correlated in the DNA and protein sequences, the evolutionary models used in phylogeny assume that these positions are evolving in an independent fashion.

Here we propose a new model that considers co-evolving positions.

The model is based on a 16 X 16 instantaneous rate matrix and three parameters: s, d, w. The parameter s is the rate associated with a transition from a co-evolving combination to a non-co-evolving one and d is the rate of a transition from one non-co-evolving combination to a co-evolving one. The additional parameter w is the rate attributed to a single mutation occurring between two non-co-evolving combinations.

To evaluate the new DNA dependent model, we use likelihood ratio test (LRT) between two models: the null model where independent evolution is assumed for each position (i.e. s=d=w) and the dependent model in which co-evolution is assumed. The power of LRT to distinguish between alternative evolutionary models is tested on simulated and empirical data.

The results show that the null model has a weaker likelihood when two positions are co-evolving, whereas in the case of independent positions, the dependent and the null models have similar likelihoods.

In the past decade several methods have been developed to identify co-evolving positions using probabilistic or combinatorial approaches.

They give a score of correlation but don't specify among the 144 possible combinations the ones that co-evolve across the phylogeny.

We used this new model to investigate for co-evolving positions and their associated set of combinations in melanochortin system. The system consists of melanocortin peptides derived from the proopiomelanocortin gene, five melanocortin receptors, two endogenous antagonists, and two ancillary proteins. Recent pharmacological and genetic studies have affirmed the role of melanocortins in pigmentation, inflammation, energy homeostasis, and sexual function. We found that these sets do not necessarily match the most frequent ones as it is usually expected.

This likelihood-based framework represents a step forward in reconstructing the evolution of co-evolving patterns based on a phylogeny with potential applications in evolutionary studies and mutagenesis experiments.

Poster 11

> **Eric Frichot** [1], Francois Mathieu [1], Guillaume Bouchard [2], Olivier Francois [1]
[1] Université Joseph Fourier Grenoble, Centre National de la Recherche, TIMC-IMAG UMR 5525, Grenoble, France
[2] Xerox Research Center Europe, Meylan, France

Least square estimates of ancestry coefficients using sparse NMF methods.

Population structure has long been recognized as being a confounding effect in genetic association studies. Estimated ancestry coefficients derived from multi-locus genotype data can be used to perform statistical correction for population stratification. It is essential for these methods to obtain accurate estimates of ancestry coefficients. In this study, we propose a machine learning approach to compute least-square estimates of individual ancestry coefficients in admixed populations. We implemented a fast algorithm using sparse non-negative matrix factorization algorithms. We applied the sparse non-negative matrix factorization algorithms to human polymorphism data sets obtained from the Human Genome Diversity Panel with number of loci ranging from a few thousands to several hundred thousand single nucleotide polymorphisms, and we compared their performance with the state-of-the-art computer program ADMIXTURE. This study provided evidence that our algorithms can compute consistent estimates of genetic admixture within run-times that are about 10-fold faster than those of ADMIXTURE.

Poster 12

> **Gabriel Leventhal** [1], Huldrych Günthard, Sebastian Bonhoeffer [1], Tanja Stadler [1]
[1] ETH Zurich, Computer Science, CAB H82.2, Universitatstr. 6, Zurich, 8092

The control, prediction and understanding of epidemiological processes requires insight into how infectious pathogens transmit in a population. The chain of transmission can in principle be reconstructed with phylogenetic methods which analyse the evolutionary history using pathogen sequence data. The quality of the reconstruction, however, crucially depends on the underlying epidemiological model used in phylogenetic inference. Until now, only simple epidemiological models have been used, which make limiting assumptions such as constant rate parameters, infinite total population size, or deterministically changing population size of infected individuals. Here we present a novel phylogenetic method to infer parameters based on a classical epidemiological model. Specifically we use the susceptible-infected (SI) model, which accounts for density-dependent transmission rates and finite total population size, leading to a stochastically changing infected population size. We first validate our method by estimating epidemic parameters for simulated data and then apply it to transmission clusters from the Swiss HIV epidemic. We show that our estimates of the basic reproductive number R0 for the considered Swiss HIV transmission clusters are significantly higher than previous estimates, which were derived assuming infinite population size. This difference in key parameter estimates highlights the importance of careful model choice when doing phylogenetic inference. In summary, this paper presents the first fully stochastic implementation of a classical epidemiological model for phylogenetic inference and thereby addresses a key aspect in ongoing efforts to merge phylogenetics and epidemiology.

Poster 13

> **Denis Fargette** [1], Pinel-Galzi [1], Ocholla [2], Rakotomalala [3]
[1] IRD Montpellier France,
[2] NARO, Kampala Uganda,
[3] FOFIFA Mahajanga Madagascar

What can be predicted about Rice yellow mottle virus emergence and evolution ?

A recent review entitled « What can be predicted about virus emergence and evolution ?» concluded that any success in predicting what may emerge is likely to be limited, but that forecasting how viruses might evolve and spread following emergence is more tractable (Holmes EC, 2012, Current Opinion in Virology, in press). This statement is tested with Rice yellow mottle virus (RYMV), a plant virus whose evolution has been thoroughly studied. We focused on the phylodynamics of RYMV and opposed two situations: that of RYMV in Madagascar, a well documented island system, which provides a simplified model in which phylogeography of the virus is tractable, to the recent emergence of a new strain around Lake Victoria which illustrates that success in predicting what may emerge in more complex situations is limited.

Poster 14

> **Maria Ines Fariello** [1,2] , S Boitard [1], H Naya [2], M San Cristobal [1], B Servin [1]
[1] Centre INRA de Toulouse Midi-Pyrénées 24 Chemin de Borde Rouge CS 52627 31326 Castanet-Tolosan cedex FRANCE
[2] Contact Institut Pasteur de Montevideo Dirección: Mataojo 2020, Montevideo CP 11400, Uruguay

Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations

The adaptation of populations to environmental changes comes with modifications of their genetic background. In particular, it is expected that genes underlying traits influenced by adaptation will exhibit larger changes than the average random variation along the genome. A major topic in modern population genetics and evolution lies in the identification of these genes. While several strategies are used for this identification, a widely used one consists in comparing related populations that have undergone different selection pressures and identifying genome regions that show outstanding differences between them.

A classical measure of the difference between populations at a single locus is Fst. While this measure has proven its efficiency, it does not account for complex relatedness between populations nor for correlation between neighbouring loci on the genome (linkage disequilibrium). We proposed a new test, called hapFLK [1], for detecting selection based on the differentiation between several populations. HapFLK extends the single SNP test FLK [2], which already accounts for the correlations between populations using the local haplotype clustering model from fastPHASE [3], to account for the LD.

The detection power of hapFLK exceeds that of FLK and Fst in various simulated scenarios. In the particular case of two population scenarios, it also exceeds that of XP-EHH [4], especially for selection on standing variation.

We applied the hapFLK and FLK tests to a set of 6 northeuropean sheep populations: our new test allowed to identify more complex genetic responses to selection than FLK (or Fst). We also proposed a method to pinpoint the population(s) under selection.

A software for computing hapFLK is available on https://forge-dga.jouy.inra.fr/projects/hapflk

[1] Fariello M.I., Boitard S., Naya H., San Cristobal M., Servin B. (2013) Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations doi:10.1534/genetics.112.147231

[2] Bonhomme M., Chevalet C., Servin B., Boitard S., Abdallah J., Blott S., San Cristobal M., Amberg, S. Romdhani, T. Vetter. (2010) Detecting selection in population trees: The Lewontin and Krakauer test extended. Genetics

[3] Scheet, P. and Stephens M. A. (2006) Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase, The American Journal of Human Genetics

[4] Sabeti P. C., Varilly P., Fry B., Lohmueller J., Hostetter E., Cotsapas C., Xie X., Byrne E. H., McCarroll S. A., Gaudet R., Schaffner S. F., and Lander E. S. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature

Poster 15

> **Mareike Fischer** [1], Leo van Iersel [2], Steven Kelk [3], Céline Scornavacca [4]
[1] Institute of Mathematics and Computer Science, Greifswald University, Germany
[2] Centrum Wiskunde & Informatica, Amsterdam University, The Netherlands
[3] Maastricht University, The Netherlands
[4] Institut des Sciences de l'Evolution de Montpellier - CC64 Université Montpellier II, France

Maximum Parsimony and Maximum Likelihood on phylogenetic networks

Ever since Darwin's first sketch of a phylogenetic tree, such trees are the model of choice for most evolutionary studies. But unfortunately, trees are unable to explain phenomena like horizontal gene transfer and hybridization. As many organisms are known to be subject to these evolutionary events, phylogenetic trees are more and more replaced by so-called phylogenetic networks. The aim is then to find the best network for a given dataset according to an optimization criterion. Two such criteria often used for phylogenetic tree reconstruction are Maximum Parsimony and Maximum Likelihood. In my talk, I will first present different ways to extend the parsimony concept from trees to networks, and I will explain some interesting properties of parsimony on networks. Moreover, I will show that in networks, even the so-called "small parsimony problem" is hard. Then, I will show how all these properties of parsimony lead to conclusions concerning Maximum Likelihood and I will show why trees play a fundamental role in finding a most parsimonious or most likely network.

Poster 16

> Thomas Berngruber, Rémy Froissart, Marc Choisy and **Sylvain Gandon**
CEFE, 1919 route de Mende, 34293 Montpellier, France

Evolution of virulence

Theory predicts that selection for pathogen virulence and horizontal transmission is highest at the onset of an epidemic but decreases thereafter, as the epidemic depletes the pool of susceptible hosts [1-4]. We tested this prediction by tracking the competition between the latent bacteriophage **λ** and its virulent mutant λcI857 throughout experimental epidemics taking place in continuous cultures of Escherichia coli. As expected, the virulent **λ**cI857 is strongly favored in the early stage of the epidemic, but loses competition with the latent virus as prevalence increases. We show that the observed transient selection for virulence and horizontal transmission can be fully explained within the framework of evolutionary epidemiology theory. This experimental validation of our predictions is a key step towards a predictive theory for the evolution of virulence in emerging infectious diseases.

Poster 17

> **Mathieu Gautier**, Renaud Vitalis
UMR INRA/CIRAD:ORD/SupAgro CBGP, Montferrier sur Lez, France

A Bayesian model to estimate the effective sex-ratio from SNP data under a pure-drift demographic model

In sexual species, evaluating the relative female and male contribution to the demography represents a mattering stake to better understand the history and the dynamics of the populations. Of particular interest, the effective sex ratio is defined as the proportion of effective number of breeding females (Nf) relative to the total effective population size (N=Nf+Nm) and provides insights into the underlying social organization. We herein present a Bayesian hierarchical model to estimate the effective sex-ratio based on an extension of a previously developed model (the K model) which aimed at estimating divergence times on a population tree (Gautier and Vitalis, 2013).

We evaluated our model via extensive simulations and compared it to previously proposed approaches (e.g. Keinan et al., 2009) that rely on standard summary statisitics of population differentatiation. We finally provide illustration on real data by analyzing different data sets from different species.

~ Gautier M and Vitalis R. (2013). Inferring population histories using genome-wide allele frequency data. Molecular Biology and Evolution. Accepted
~ Keinan A, Mullikin JC, Patterson N, Reich D (2009) Accelerated genetic drift on chromosome x during the human dispersal out of africa. Nat Genet 41: 66-70.

Poster 18

> **Erida Gjini** [1,2], Daniel T. Haydon [2,3,4], J. David Barry [4], Christina A. Cobbold [1,2]
[1] School of Mathematics and Statistics, College of Science and Engineering, University of Glasgow, Glasgow, United Kingdom
[2] The Boyd Orr Centre for Population and Ecosystem Health, University of Glasgow, Glasgow, United Kingdom
[3] Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary, and Life Sciences, University of Glasgow, Glasgow, United Kingdom
[4] Wellcome Trust Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, University of Glasgow, Glasgow, United Kingdom

Modelling local diversification in antigen gene families: an example from African trypanosomes

Patterns of genetic diversity in parasite antigen gene families hold important information about their potential to generate antigenic variation within and between hosts. The evolution of such gene families is typically driven by gene duplication, followed by point mutation and gene conversion. There is great interest in estimating the rates of these processes from molecular sequences for understanding the evolution of the pathogen and its significance for infection processes. In a recent study (1), we construct a series of models to investigate hypotheses about the nucleotide diversity patterns between closely related gene sequences from the antigen gene archive of the African trypanosome, the sleeping sickness parasite. Using a hidden Markov model, we identify two scales of diversification: clustering of sequence mismatches, a putative indicator of gene conversion events with other lower-identity donor genes in the archive, and at a sparser scale, isolated mismatches, likely arising from independent point mutations. In addition to quantifying the respective probabilities of occurrence of these two processes within a Bayesian framework, we also obtain estimates for the gene conversion tract length distribution and the average diversity contributed locally by conversion events.

1) Gjini E, Haydon DT, Barry JD, Cobbold CA.(2012) The impact of mutation and gene conversion on the local diversification of antigen genes in African trypanosomes, Mol Biol Evol. 29(11):3321-31.

Poster 19

> **Matthew Hall** [1], Ruth Zadoks [2], Andrew Rambaut [1], Mark Woolhouse [3]
[1] Institute of Evolutionary Biology, Ashworth Laboratories, University of Edinburgh, Edinburgh, United Kingdom
[2] Institute of Biodiversity Animal Health and Comparative Medicine, University of Glasgow, Glasgow, United Kingdom

[3] Centre for Immunity, Infection and Evolution, Ashworth Laboratories, University of Edinburgh, Edinburgh, United Kingdom

Co-estimation of phylogenetic and transmission trees for infectious disease outbreaks

The reconstruction of infectious disease transmission networks from genetic data has been the subject of a number of recent papers, particularly in the field of animal disease (e.g. [1-3]). In those published so far, the genetic relationships between isolates have been modelled in relatively simple terms, using either measures of pairwise genetic distance or trees constructed using statistical parsimony. This poster describes a method currently in development to integrate analysis of this sort with the full Bayesian phylogenetics framework available in BEAST [4], allowing the insights of current phylogenetic methods, such as estimation of the dates of ancestral nodes in the phylogenetic tree, to be brought to bear on the problem. It is based on the assignment of an infected unit to each internal node of the phylogenetic tree to represent the location or individual in which the respective ancestor existed, and the observation that if no reinfection occurs, the set of nodes that are associated with this unit must form a connected subgraph of the tree. This provides a means to overlay the transmission tree onto the phylogenetic tree. Monte Carlo Markov Chain methods are used to sample from the joint distributions of both, with the likelihood of the transmission tree being calculated using epidemiological case data.

The method enables the estimation of epidemiological parameters, such as latent periods, and measures of spatial spread. Examination of the posterior set of transmission trees also gives insight into the properties of the network and the reproduction numbers of the epidemic. It also potentially allows for uncertainty in the phylogeny and parameters of the evolutionary model to be reduced by inclusion of epidemiological data.

[1] EM Cottam, et al. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus". P Roy Soc B-Biol Sci 275.1637 (2008), pp. 887-895.
[2] MJ Morelli, et al. A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data". PLoS Comput Biol 8.11 (2012), e1002768.
[3] RJF Ypma, et al. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data". P Roy Soc B-Biol Sci 279.1728 (2011), pp. 444-450.
[4] AJ Drummond, MA Suchard, D Xie, and A Rambaut. Bayesian Phylogenetics with BEAUti and the BEAST 1.7". Mol Biol Evol 29.8 (2012), pp. 1969-1973.

Poster 20

> **Denise Kühnert** [1,2], Alexei Drummond [1,2]
[1] The University of Auckland
[2] Allan Wilson Centre, New Zealand

Understanding virus epidemics through analysis of genomic data

Ecological and evolutionary processes governing rapidly evolving viruses act on the same timescale: The evolution of such viruses (namely HIV, HCV and Influenza virus etc.) is closely entangled with the host population dynamics. The cross-reaction of the two processes must be accounted for when inferring epidemiological parameters and/or phylogenetic history.

Our aim is a joint epidemiological phylogeographic analysis of genomic data by incorporating the dynamics of a geographic Susceptible-Infected-Recovered model (SIR) into Bayesian phylogenetic inference.

In order to achieve this, we first extend the serial birth-death model [1], which allows birth, death and sampling rates to change over time [2]. This extension not only reconstructs epidemiological parameters such as the basic reproduction number, but it can also detect changes in those parameters over time. Such changes reflect on the efficiency of measures such as prophylaxis or treatment campaigns.

The possibility of rate changes over time enables coupling of the birth death process with an SIR process. Therefore, the next step is the development of a birth death SIR model (BDSIR) in a single population: In a Bayesian MCMC framework, the epidemiological process is forward simulated such that it can be used to approximate the per lineage rates of a birth-death-sampling process on the number of infected individuals in an epidemic. When the relevant rates change, the SIR simulation is being updated using either a Step Anticipation tau-Leaping approach [3]. Thus, incidence and prevalence are part of the methods output.

Finally, we assume underlying structured populations with migration between discrete locations. A further extension of the birth-death model allows for migration as part of the process as well as differing rates among different locations. This enables a structural BDSIR model in which every discrete location is considered as a single epidemic that can be caused and affected by a migration event from another infected population.

References
[1] Stadler T. et al. Estimating the basic reproductive number from viral sequence data. Molecular biology and evolution, 2012
[2] Stadler T., Kühnert D. et al. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and HCV. PNAS, in press)
[3] Li T. Analysis of explicit tau-leaping schemes for simulating chemically reacting systems, Multi. Mod. Simul. 6, 2007

Poster 21

> **Raphaël Leblois**, Beeravolu, Rousset, Pudlo
Centre de Biologie et Gestion des Populations (CBGP, UMR 1062), France

Likelihood-based inferences from genetic data under simple epidemiological scenarios

The inference of past changes in population size from classical population genetic samples (i.e. non-genomic data) usually relies on relatively simple models (i.e. with a single discrete or continuous demographic change). With the help of these models, we can detect past contractions/expansions, infer the past and present population sizes and the time at which these events occurred. In the case of epidemiological events, demographic histories often consist of a founder event followed by an expansion. In order to make demographic inference under such "Founder-Flush" scenarios, we pursued a likelihood-based method. The likelihood algorithm is based on the importance sampling scheme of Stephens and Donnelly (2000) and De Iorio and Griffiths (2004a,b). We tested the efficiency of our approach using simulations. More specifically, we studied the precision of the model parameter estimates with respect to the quantity and type of genetic makers (microsatellites, DNA sequences and Single Nucleotide Polymorphisms, SNPs). We also attempted to increase the precision of the estimates by implementing an inference which simultaneously handles different types of markers in a single analysis.

Poster 22

> **Thi Hau Nguyen** [1,2], Vincent Ranwez [2], Celine Scornavacca [3] and Vincent Berry [1,4]
[1] LIRMM, University Montpellier 2 - CNRS, France
[2] Montpellier SupAgro (UMR AGAP), France
[3] ISEM, UMR 5554, University Montpellier 2, France
[4] Institut de Biologie Computationnelle, LIRMM, UM2, CNRS, Montpellier, France

Inferring confidence values of evolutionary events

Reconciliation methods compare gene trees and species trees to recover evolutionary events explaining the history and composition of genomes, such as duplications (D), transfers (T) and losses (L). These methods play an important role in studying genome evolution as well as in inferring orthology relationships. A major issue with reconciliation methods is that the reliability of the predicted evolutionary events may be put into question due to a number of reasons: Firstly, there possibly exists multiple equally optimal reconciliations for a given pair of species tree and gene tree. Secondly, reconciliation methods can be misled by inaccurate gene or species trees [Hahn 2007]. Thirdly, the predicted events vary as the method parameters — such as the costs of the elementary events — change. For all these reasons, confidence values for the predicted evolutionary events are deeply needed. In this work, we revise and enhance the method of [Scornavacca et al. 2012] that uses a reconciliation graph to infer such confidence values based on their frequencies in a set of reconciliations. The original method considers the set of equally parsimonious reconciliations. We study here alternative reconciliation sets obtained by: i) computing parsimonious reconciliations for slightly different costs of the elementary events (D, T, L) or ii) probabilistically sampling suboptimal reconciliations from the dynamic cost matrix. Experiments on simulated data show the meaningfulness of event supports provided by our methods. Indeed, the evolutionary histories composed of events with high supports (≥ 50%) have higher accuracy than the ones proposed by the traditional reconciliation tools, which do not use supports.

Poster 23

> **Duncan Palmer**, Angela McLean, Gil McVean
Department of Statistics, South Parks Road, University of Oxford
Department of Zoology, South Parks Road, University of Oxford

Integrating genealogical and dynamical modelling to infer escape and reversion rates in HIV epitopes

The rates of escape and reversion in response to selection pressure arising from the host immune system, notably the cytotoxic T-lymphocyte (CTL) response, are key factors determining the

evolution of HIV. Existing methods for estimating these parameters from cross-sectional population data using ordinary differential equations (ODE) ignore information about the genealogy of sampled HIV sequences, which has the potential to cause systematic bias and over-estimate certainty. We describe an integrated approach, validated through extensive simulations, which combines genealogical inference and epidemiological modelling, to estimate rates of CTL escape and reversion in HIV epitopes. We show that there is substantial uncertainty about rates of viral escape and reversion from cross-sectional data, which arises from the inherent stochasticity in the evolutionary process. By application to empirical data, we find that point estimates of rates from a previously published ODE model and the integrated approach presented here are often similar, but can also differ several-fold depending on the structure of the genealogy. The model-based approach we apply provides a framework for the statistical analysis of escape and reversion in population data and highlights the need for longitudinal and denser cross-sectional sampling to enable accurate estimate of these key parameters.

Poster 24

> **Bárbara Parreira**, Isabel Gordo, Lounès Chikhi
Instituto Gulbenkian de Ciência. Rua da Quinta Grande, 6, 2780-156 Oeiras, Portugal
CNRS, EDB (Laboratoire Evolution et Diversité Biologique), UMR CNRS/UPS 5174, F-31062 Toulouse, France

The genetic consequences of fine scale structure: social groups and intra-host structure as little studied examples

Most organisms live in subdivided populations where individuals interact locally. It is well recognized that this fact has biological consequences. Indeed, theoretical studies predict ecological dynamics and genetic diversity differences between structured populations and nonsubdivided/ homogeneous populations. However, while some levels of structure are easily recognized and studied, others are often neglected. At the population level, it is well accepted that the genetic diversity is dependent on processes such as migration and dispersal, habitat fragmentation, habitat selection or species spatial distribution. This can for instance result in an increase in drift within subpopulations or local adaptation promoted by different selection pressures.

However, it is at the level of the population or the host (in a host-parasite system) that structure is usually ignored. Indeed, in many species individuals are organized in social structures, which are usually ignored by population geneticists, even though this fine scale structures have been shown to have consequences on patterns of genetic diversity (Nunney 1999, Sugg et al. 1996, Chesser et al. 1993). For instance, breeding tactics do have consequences in the variance of the reproductive success, local inbreeding or gene correlations. Also, in epidemiological studies several studies have demonstrated that intra-host structure plays a role in the diversification and evolution of pathogens (Shriner 2006, Frost 2001) but most studies tend to consider the host as a panmictic unit.

With the increasing availability of computational and simulation tools it is important to now integrate the different levels of population structure present in natural systems. Indeed, the manner in which these fine scale structures affect population genetic properties is still not well understood.

During my PhD I have been interested in two different types of fine-scale structures: social structure and intra-host structure in order to understand how they shape the distribution of diversity within populations. We used published data on the social structure of several vertebrate species and on intra-host virus data and computed commonly used population genetics statistics.

In order to quantify how genetic variability is influenced by the smaller level structures (i.e. the existence of social groups within populations or pathogens within hosts) we performed computer simulations under different subpopulation connectivity topologies. Results suggest that genetic diversity obtained in real data can be better reproduced when these fine scale structures are taken into account in the modeling approach.

Sugg H. W., et al. (1996) doi10.1016/0169-5347(96)20050-3
Chesser R. K. (1993) Influence of gene flow and breeding tactics on gene diversity within population. Genetics.129: 573-583
Frost S.D. et al. (2001) doi10.1073/pnas.131056998
Nunney L. (1999) doi10.2307/2640915
Shriner D. et al. (2006) doi10.1554/05-473.1

Poster 25

> **Andrei-Alin Popescu**, Katharina T. Huber
School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK

Lassoing and corraling rooted phylogenetic trees

The construction of a dendogram on a set of individuals is a key component of, for example, genomewide association studies. However even with modern sequencing technologies the distances on the individuals required for the construction of such a structure may not always be reliable making it tempting to exclude them from an analysis. This, in turn, results in an input set for dendogram construction that consists of only partial distance information which raises the following fundamental question. Suppose we have a dendogram T with leafset X. Then for what sets L of pairs of elements of X, is T uniquely determined by the distances T induces on the elements of L?

In this talk, we first formalize the notion of a dendogram in terms of a certain edgeweighted rooted phylogenetic tree and then study four different interpretations of the idea of uniquely determining giving rise to four distinct types of lassos.

Poster 26

> **Francois Rousset** [1,2], Jean-Baptiste Ferdy[3]
[1]Université Montpellier 2, CNRS, Institut des Sciences de l'Evolution, France
[2]Institut de Biologie Computationnelle, Montpellier
[3]Laboratoire Evolution et Diversité Biologique, Université Toulouse 3

Testing environmental and genetic effects in correlated landscapes

Correlated random effects are a well-recognized concern for observational data in general, and more specifically for spatial data in ecology. Generalized linear mixed models (GLMMs) with correlated random effects are a potential framework for handling these correlations. However, as the result of statistical and practical issues, such GLMMs have been fitted through the undocumented use of procedures based on penalized quasi-likelihood approximations (PQL), and under restrictive models of spatial correlation. Alternatively, they are often neglected in favor of simpler but more questionable approaches such as partial Mantel tests. In this work we provide practical and validated means of inference under spatial GLMMs, that overcome these limitations. For this purpose, an R package has been developed to fit spatial GLMMs. We assess the performance of likelihood ratio tests for fixed effects under spatial autocorrelation, based on Laplace or PQL approximations of the likelihood. Expectedly, the Laplace approximation performs generally slightly better than PQL, although a variant of PQL was better in the binary case, where it could be expected from previous works to perform worst. Finally, we illustrate the efficiency of a bootstrap procedure for correcting the small sample bias of the tests, which should be of more general interest for non-spatial models. We therefore provide an implementation of spatial GLMMs, suitable for testing environmental and genetic effects in small samples.

Poster 27

> **Sandoval-Castellanos Edson**
Department of Molecular Systematics, Swedish Museum of Natural History, Svante Arrhenius väg 9, 114 18, Stockholm

Direct Bayesian Simulation of Evolutionary Trees

Computer based methods are becoming increasingly popular for statistical inference in evolutionary genetics studies. They are aimed to treat large and complex systems. However, there is a widespread concern that those analysis are increasingly been constrained by the available computational power, and the advent of technologies as Next-generation sequencing could sensibly worsen the problem.

Evolutionary trees in their different forms (phylogenetic, genealogical or coalescent) are essential to most of theory and statistical inference of evolutionary genetics. Their simulation from a probability distribution represent an essential part when not the whole of the calculation procedure performed for statistical inference in most of Monte Carlo techniques applied to evolutionary genetics studies. Efficient ways of simulating such trees could improve those Monte Carlo techniques.

There are two types of algorithms for simulating evolutionary trees either for Monte Carlo simulation or for a final estimation: sequential algorithms or local search algorithms. Sequential algorithms lack of relevant statistical properties for simulation since they employ non probabilistic criteria. On the other hand local search algorithms have problems with finding optimal search strategies and thus are computationally costly. This dichotomy is boosting the development of hybrid and parallelizable methods.

Here, I present a third way to simulate evolutionary trees by simulating a tree by incorporating the information contained in the observed data (a DNA alignment) along with the stochasticity provided by the unknown information and the properties of the adopted model. The algorithm retrieves the information contained in a DNA alignment to simulate first the topology and afterwards the branches lengths. For that purpose, it is required a detailed analysis of the model to fathom the flow of information from the tree into the data in order to build an algorithm that retrieves randomly such information back to a tree from a given dataset. That procedure allows skipping the calculation of likelihoods but still obtain the trees according with their posterior probabilities. Moreover, if the procedure is intended for integration purposes, it could enable the marginalization of the densities of parameters of interest (conditional to the observed data), in order to estimate those parameters. In addition, hypothesis contrasts can be performed by classifying either the simulated trees or the parameters simulated.

The efficiency of this method makes it an interesting alternative for analyzing Next-generation sequencing data with a limited computational budget.

Poster 28

> **Elke Schaper** [1], Maria Anisimova [2], Olivier Gascuel [3,4]
[1] CNB H107.2 Universitätstr. 6, CH-8092 Zürich
[2] CAB H82.2.2 Universitätstr. 6, CH-8092 Zürich
[3] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM), UMR5506 CNRS, Université Montpellier 2, France;
[4] Institut de Biologie Computationnelle (IBC), 95 Rue de la Galéra, 34095 Montpellier, France

The evolution of protein tandem repeats

Tandem repeats represent one of the most prevalent features of proteomic sequences. However, little is known about the mechanisms and time-scales of tandem repeat formation nor the functional relevance of most tandem repeats. Here, we would like to present recent results on the evolution of tandem repeats in three parts.

First, we show a benchmark of current tandem repeat detection algorithms on simulated and real sequence data. Tandem repeat detection accuracy, power and prediction quality vary strongly between tandem repeat detectors. Furthermore, the detection power depends heavily on the tandem repeat unit length, the number of repeat units, and the divergence of the tandem repeat (1).

Second, we introduce a model-based statistical test to distinguish between true and false tandem repeat detections, and in this way controlling the false positive detection rate. In short, we distinguish between the null hypothesis that alleged tandem repeat units are not related, and the alternative hypothesis, that they evolved from a common ancestral repeat unit by duplication (1).

Third, we show a representation of tandem repeats by hidden Markov models useful for the detection of homologous tandem repeats. Comparing homologous tandem repeats across the Eucaryotic clade, the conservation of tandem repeats across this evolutionary time scale can be analysed. We show to what extent the current assembly of tandem repeat units is still informative about the tandem repeat phylogeny, and examine how substitution rates vary across tandem repeat containing proteins. Lastly, we analyse how tandem repeat formation events are distributed over this time scale.

1. Schaper,E., Kajava,A.V., Hauser,A. and Anisimova,M. (2012) Repeat or not repeat?--Statistical validation of tandem repeat prediction in genomic sequences. Nucleic Acids Research, 40.

Poster 29

> **Adam Szalkowski**, Maria Anisimova
Universitätstrasse 6, 8092 Zürich, Switzerland

Graph-based approach to improve global alignment of sequences with tandem repeats

Many proteins with crucial biological functions contain tandem repeats (TRs). TRs are present in many proteins responsible for resistance or pathogeneicity, and those associated with several infectious and neurodegenerative diseases. This motivates numerous studies of TRs and their evolution, requiring accurate multiple sequence alignment. Due to replication slippage, TR units may be lost or inserted at any position of a TR region. Despite this, current alignment methods assume fixed domain boundaries, but are still of high complexity. Here we present a new global alignment method that allows for TR mutations which are not restricted by TR unit boundaries. The method is implemented in ProGraphMSA+TR, an extension of a fast phylogeny-aware graph-based method ProGraphMSA, enhanced with a correction for presence of TRs. As the TR indels are modeled separately, our implementation unlike any previous tools, enables the rate estimation for TR unit indels. As a result our method allows to reconstruct accurate global alignments, disentangling TR units and measuring TR unit indel events in a biologically meaningful way. Our method is not constrained to duplication events and is able to detect all changes in TR regions due to recombination, strand slippage, and other events which insert or delete one or multiple TR units. We evaluate our algorithm by simulation incorporating TR evolution, by either sampling TR units from a profile hidden Markov model or by mimicking strand slippage with duplications. The utility of the new method is illustrated on the GALA-LRR proteins, a family of type III effectors and the major pathogenicity determinant of the agriculturally relevant Ralstonia solancefarum bacteria species. We show that TR indel rate variation makes an important contribution to the diversification of GALA-LRR subfamilies.

Poster 30

> **Gergely J. Szöllősi** [1], Wojciech Rosikiewicz [2], Bastien Boussau [1,3], Eric Tannier [1] and Vincent Daubin [1]

[1] LBBE, CNRS, UMR 5558, Université Lyon 1, F-69622 Villeurbanne, France;

[2] Laboratory of Bioinformatics, Faculty of Biology, Adam Mickiewicz University, Poznán, Poland;

[3] Department of Integrative Biology, University of California, Berkeley, California, United States of America.

Efficient Exploration of the Space of Reconciled Gene Tree

As molecular phylogeneticists, we infer gene trees based on sequence information. Unfortunately, sequences alone contain limited signal, and as a result phylogenetic reconstruction almost always involves choosing between statistically equivalent or weakly distinguishable relationships. Although each homologous gene family has its own unique story, they are all related by a shared species history, which could be helpful for gene tree inference. We have recently published a probabilistic reconciliation model, which describes the relationships between a gene tree and a species tree as a series of events, such as duplication, transfer and loss, speciation and extinction (Szöllősi et al. Syst. Biol. 2013). We now propose an efficient way to integrate sequences and reconciliation information in the inference of gene trees.

To design a species tree aware method for reconstructing gene phylogenies, the space of reconciled gene trees must be explored using information from both a model of sequence evolution and a reconciliation model. Such an exploration can be tedious with classical approaches. To circumvent this problem, we present a general probabilistic approach to exhaustively explore all reconciled gene trees that can be amalgamated as a combination of clades observed in a sample of gene trees. For a sample derived from the posterior distribution of trees obtained from a bayesian MCMC analysis, this approach provides an accurate approximation of gene tree likelihood.

We demonstrate using both simulations and biological sequences that gene phylogenies reconstructed using the joint likelihood are dramatically more accurate than those reconstructed using sequences alone. In fact, we find that even using a simplistic model of sequence evolution, the joint reconstruction yields significantly more accurate gene trees than the sequence-based inference with the complex model used in simulations. Considering 1099 homologous gene famillies from 36 genomes of cyanobacteria we find that the majority of phylogenetic discord results from errors in sequence based reconstruction that can be corrected using information aggregated across gene families by a putative species tree. The result is a striking reduction in apparent phylogenetic discord, with resp. 24%,$59% and 46% percent reductions in the mean numbers of duplications, transfers and losses per gene family.

Our probabilistic method overcomes a fundamental limitation of recent parsimony based methods to improve gene trees given a putative species tree (David and Alm Nature 2011,Wu et al. Syst. Biol. 2013) by not having to rely on any ad hoc assumption about statistical support, while at the same time deploying approximations that make it more efficient than methods that rely on a local search of tree space (Akerborg et al. PNAS 2009).

The open source implementation of the method is available from https://github.com/ssolo/ALE.git .

References:

~ Akerborg, O., B. Sennblad, L. Arvestad, and J. Lagergren. 2009. Simultaneous bayesian gene tree reconstruction and reconciliation analysis. Proc Natl Acad Sci U S A 106:57149.

~ David, L. A. and E. J. Alm. 2011. Rapid evolutionary innovation during an archaean genetic expansion. Nature 469:93 6.

~ Szöllősi, G. J., E. Tannier, N. Lartillot, and V. Daubin. 2013. Lateral gene transfer from the dead. Systematic Biology doi: 10.1093/sysbio/syt003

~ Szöllősi G. J., Rosikewicz W., Boussau B., Tannier E. and Daubin V. Efficient Exploration of the space of reconciled trees Systematic Biology (under review)

~ Wu, Y.-C., M. D. Rasmussen, M. S. Bansal, and M. Kellis. 2013. Treefix: Statistically informed gene tree error correction using species trees. Systematic Biology 62:11020.

Poster 31

> **Leo van Iersel**, Steven Kelk, Nela Lekić , Leen Stougie
Science Park 123

Approximation algorithms for nonbinary agreement forests

Given two rooted phylogenetic trees on the same set of taxa X, the Maximum Agreement Forest problem (MAF) asks to find a forest that is, in a certain sense, common to both trees and has a minimum number of components. The Maximum Acyclic Agreement Forest problem (MAAF) has the additional restriction that the components of the forest cannot have conflicting ancestral relations in the input trees. These problems have been studied intensively because of their close relationship to, respectively, the rooted Subtree Prune and Regraft distance and the Hybridization Number of the input trees. Almost all these studies restrict their attention to the binary case. However, in practice, phylogenetic trees are rarely binary due to uncertainty about the precise order of speciation events. Here we show how approximation algorithms can be obtained for the general, nonbinary variants of these problems.

Poster 32

> **Renaud Vitalis** [1], Gautier M. [1], Dawson K.J. [2], Beaumont M.A. [3]
[1] INRA, UMR CBGP (INRA, IRD, CIRAD, Montpellier SupAgro), Campus International de Baillarguet, 34988 Montferrier-sur-Lez cedex, France
[2] Cancer Genome Project, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK
[3] Department of Mathematics and School of Biological Sciences, University of Bristol, Bristol BS8 1TNW, UK

Detecting and measuring selection from gene frequency data

How best to identify regions, loci or single nucleotides that have been, or still are, under selection is still a challenging issue. Here, we provide a new method to distinguish neutral from selected polymorphisms and estimate the intensity of selection at the latter. This method is based on a diffusion approximation for the distribution of allele frequency in a population subdivided in a number of demes that exchange migrants. The framework for statistical inference from this model consists in a hierarchical Bayesian model. We use a Monte Carlo Markov Chain (MCMC) approach for sampling from the joint posterior distribution of the model parameters. We evaluate the statistical performance of the method using simulated data, and show how the method performs using (real) human data.

Poster 33

> **Melissa J. Ward** [1], Samantha J. Lycett [1], Marcia L. Kalish [1], Andrew Rambaut [1,3], Andrew J. Leigh Brown [1],
[1] University of Edinburgh, Institute of Evolutionary Biology, Ashworth Laboratories, Edinburgh, United Kingdom
[2] Vanderbilt University, Vanderbilt Institute for Global Health, Nashville, Tennessee, USA
[3] Fogarty International Center, National Institutes of Health, Bethesda, Maryland, USA

Estimating recombination rates from phylogenies

Understanding recombination as an ancestral process is important for unravelling the evolutionary history of viruses such as HIV and explaining observed patterns of viral diversity. Furthermore, failure to account for recombination can confound phylogenetic inference and analyses of selection. Recombination can be detected along a sequence alignment because it creates phylogenetic discordance, and this property can be exploited to estimate the rate at which recombination has occurred across the history of a population. I present a phylogenetic method [1] for quantifying recombination from genetic sequence data, using discrete ancestral state mapping methods implemented in BEAST [2,3,4]. The method is applied to sequence data from early HIV-1 group M in Kinshasa, the capital of the Democratic Republic of Congo. Kinshasa has been implicated as the epicentre of the HIV-1 group M epidemic and is unique in that almost all of the subtypes A-K circulate there. Previous analysis has revealed that isolates from a number of patients fall in different positions in phylogenetic trees constructed from sequences from opposite ends of the genome, as a result of recombination between viruses of different subtypes [5]. I obtain an estimate of the rate at which inter-subtype recombination has contributed to the observed diversity of HIV-1 group M in Kinshasa and discuss its evolutionary and epidemiological implications. I describe how the method could be used to elucidate restrictions to inter-subtype recombination which have been suggested by in-vitro studies of HIV-1 group M, and may also be applied to other viral datasets, for example to investigate rates of reassortment between influenza segments. I also consider the relative merits of different discrete trait mapping approaches for such purposes.

References:
1. Ward MJ, Lycett SJ, Kalish ML, Rambaut A, Leigh Brown AJ (2013). Estimating the Rate of Recombination in Early HIV-1 Group M Strains. J. Virol. 87(4): 1967-1973.
2. Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009). Bayesian Phylogeography Finds Its Roots. PLoS. Comput. Biol. 5: e1000520.
3. Minin VN, Suchard MA (2008). Counting labeled transitions in continuous-time Markov models of evolution. J. Math. Biol. 56: 391-412.
4. Minin VN, Suchard MA (2008). Fast, accurate and simulation-free stochastic mapping. Phil. Trans. R. Soc. London B Biol. Sci. 363: 3985-3995.
5. Kalish ML, Robbins KE, Pieniazek D, Schaefer A, Nzilambi N, et al. (2004). Recombinant viruses and early global HIV-1 epidemic. Emerg. Infect. Dis. 10: 1227-1234.

Poster 34

> **Rolf Ypma** [1], WM van Ballegooijen [1], J Wallinga [1]
[1] RIVM, Antonie van Leeuwenhoeklaan 9, 3721 MA Bilthoven, the Netherlands

Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data

Traditionally, unraveling the disease dynamics for a given outbreak has been done by using epidemiological data (e.g. day of symptom onset, length of illness). Recently, much attention has gone into incorporating genetic data of the pathogen into these analyses [1-4]. In such analyses, possible transmission trees are evaluated based on the available genetic and epidemiological data. This results in a probabilistic transmission tree, which gives the probability for any possible infection event.

Current methods rely on a number of simplifications and approximations; the genetic and epidemiological are not considered simultaneously [1], not consistently [2], independence is wrongly assumed [3] or transmission times are equated with coalescent times [4]. Here, we aim to solve these shortcomings by borrowing ideas from recent research in phylodynamics [5,6]. In this field, the phylogenetic tree resulting from genetic sequences is estimated together with a specific underlying epidemiological model. When considering outbreaks, the epidemiological model is equivalent to the full set of transmissions. This means we have to estimate both the transmission tree and the phylogenetic tree conditioned on this transmission tree. We show how the keystone needed to correctly couple these two trees together is given by a within host model of the pathogen.

By bringing transmission tree reconstruction into the field of phylodynamics, a rich statistical toolbox becomes available. We can use advanced phylogenetic inference techniques, and couple these to our understanding of transmission mechanics. In our view, this constitutes the next logical step in the field of probabilistic transmission tree reconstruction.

1. Cottam EM, Thebaud G, Wadsworth J, et al. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. Proc Biol Sci 2008;275(1637):887-95.
2. Jombart T, Eggo RM, Dodd PJ, Balloux F. Reconstructing disease outbreaks from genetic data: a graph approach. Heredity (Edinb) 2011;106(2):383-90.
3. Ypma RJ, Bataille AM, Stegeman A, et al. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proc Biol Sci 2012;279(1728):444-50.
4. Morelli MJ, Thebaud G, Chadoeuf J, et al. A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. PloS Comp Biol 2012;8(11):e1002768
5. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SD. Phylodynamics of infectious disease epidemics. Genetics 2009;183(4):1421-30.
6. Stadler T. Sampling-through-time in birth-death trees. J Theor Biol 2010;267(3):396-404.

ATTENDEES

Alizon	Samuel	Laboratoire MIVEGEC
Anisimova	Maria	ETH Zurich
Arndt	Peter	Max Planck Institute for Molecular Genetics, Berlin, Germany
Bacak	Miroslav	Max Planck Institute for Mathematics in the Sciences
Bastkowski	Sarah	University of East Anglia
Bedford	Trevor	University of Edinburgh
Beerenwinkel	Niko	ETH Zürich
Benner	Philipp	Max Planck Institute for Mathematics in the Sciences
Berry	Vincent	Institut de Biologie Computationnelle, LIRMM, UM2, CNRS, Montpellier
Bielejec	Filip	KU Leuven
Binet	Manuel	Institut de Biologie Computationnelle, LIRMM, UM2, CNRS, Montpellier
Blum	Michael	CNRS, UJF Grenoble
Boenhoeffer	Sebastian	ETH Zürich
Boussau	Bastien	University of California
Champak	Beeravolu Reddy	INRA
Chan	Yao-ban	ISEM, Université Montpellier 2
Chifolleau	Anne-Muriel	Institut de Biologie Computationnelle, LIRMM, UM2, CNRS, Montpellier
Chor	Benny	School of Computer Science, Tel-Aviv University
Cornillot	Emmanuel	Université Montpellier I
Davila	Miraine	UPMC
De Maio	Nicola	Institute of Population Genetics, VetMedUni Vienna
Delsuc	Frédéric	Institut des Sciences de l'Evolution, CNRS-IRD-Université Montpellier 2
Dib	Linda	Université de Lausanne
Drummond	Alexei	University of Auckland
E Leventhal	Gabriel	ETH Zürich
Fargette	Denis	IRD
Fariello	Maria Ines	INRA
Fischer	Mareike	University of Greifswald
Frichot	Eric	TIMC-IMAG
Gandon	Sylvain	CNRS
Gascuel	Olivier	Institut de Biologie Computationnelle, LIRMM, UM2, CNRS, Montpellier
Gautier	mathieu	INRA
Gjini	Erida	Instituto Gulbenkian de Ciencia
Guénoche	Alain	IML - CNRS
Hall	Matthew	University of Edinbugh
Hartfield	Matthew	IRD Montpellier
Hiller	Michael	Max Planck Institute for Molecular Cell Biology and Genetics
Holmes	Ian	University of California
Kelk	Steven	University of Maastricht

Kühnert	Denise	The University of Auckland
Lambert	Amaury	UPMC Univ Paris 06 and Collège de France
Leblois	Raphael	INRA
Lebre	Sophie	Universite de Strasbourg
Lefort	Vincent	Institut de Biologie Computationnelle, LIRMM, UM2, CNRS, Montpellier
Leviyang	Sivan	Georgetown University
Lycett	Samantha	University of Edinburgh
Martin	Darren	Institute of Infectious Diseases and Molecular Medicine, University of Cape Town
Matsen	Erick	Fred Hutchinson Cancer Research Center
McVean	Gil	University of Oxford
Merle	Coralie	Université Paris-Sud
Nguyen	Thi Hau	University of Montpellier 2
Numminen	Elina	University of Helsinki
Palmer	Duncan	University of Oxford
Pardi	Fabio	Institut de Biologie Computationnelle, LIRMM, UM2, CNRS, Montpellier
Parks	Sarah	EBI
Parreira	Bárbara	Instituto Gulbenkian de Ciência
Peischl	Stephan	University of Bern
Popescu	Andrei-Alin	University of East Anglia
Rousset	Francois	CNRS
Sandoval Castellanos	Edson	Swedish Museum of Natural History
Schaper	Elke	ETH Zürich
Schwarz	Roland	University of Cambridge / EBI
Scornavacca	Celine	CNRS
Shirreff	George	ETH
Stadler	Tanja	ETH Zürich
Szalkowski	Adam	ETH Zürich
Szollosi	Gergely	LBBE
Van Iersel	Leo	Centrum Wiskunde & Informatica (CWI)
Vitalis	Renaud	INRA
Ward	Melissa	University of Edinburgh
Ypma	Rolf	National Institute of Publich Health and the Environment