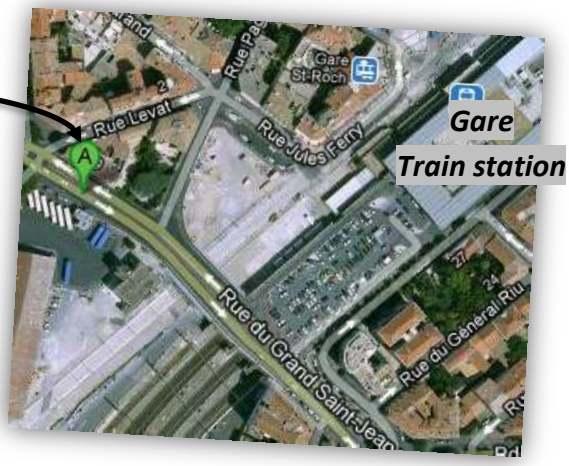
MATHEMATICAL AND COMPUTATIONAL EVOLUTIONARY BIOLOGY 2012

- INFORMATIONS-

Meeting point

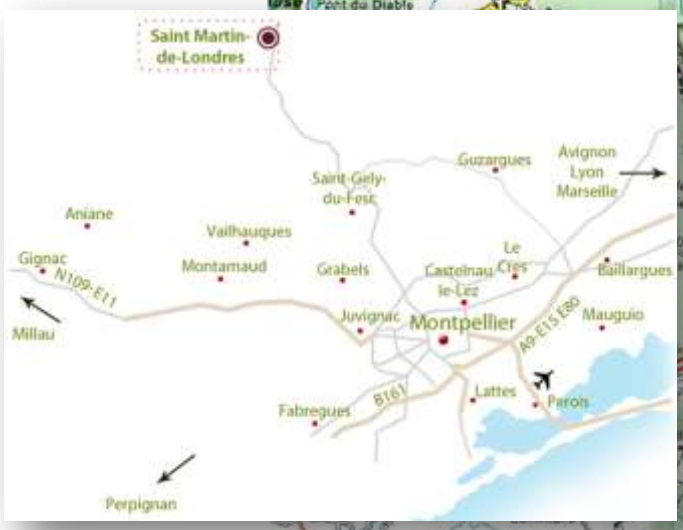
Bus station, Parking du grand Saint-Jean
(Close to the train station and server by the tram).



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).

Coordonnées : 43°49'7.80" Nord / 3°41'56.01" Est
Indication sur Google Earth, GPS, Mappy, ViaMichelin :
"Le Cayla, Saint-Martin-de-Londres"



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
-> Réductions de -20% pour les clients du hameau
Contactez donc Bernard en priorité au **06 81 16 93 75**
Tarifs : En semaine, comptez : Station occitanie du Tram = 45 € / Gare =55 € / Aéroport=65 €
Week - end / Nuit : + 15 euros

Suggested 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

MATHEMATICAL AND COMPUTATIONAL EVOLUTIONARY BIOLOGY 2012

- PROGRAM -

MONDAY 18

09:00	Bus from Montpellier to Hameau de l'Etoile
10:30	Café & croissants
10:50	Bienvenue
11:00 – 12:30	V.MOULTON - Recent progress on phylogenetic networks.
12:45	Déjeuner
14:30 – 16:00	A.ESTOUP - ABC (Approximate Bayesian Computation) methods to make inference about population history from molecular data: principles and applications.
16:00	Thé & gateaux
16:30 – 18:10 (20MIN)	D.BRYANT - Efficient lying with simulations. T.STADLER - How much do species limit each other? S.WHELAN - The effect of multiple sequence alignment methodology on downstream evolutionary analyses. C.SCORNAVACCA - Constructing minimal phylogenetic networks from softwired clusters is fixed parameter tractable. M.FISCHER - Phylogenetically decisive taxon coverage.
19:30	Apéritif
20:00	Dîner

TUESDAY 19

09:15 – 10:45	N.ROSENBERG **-** Models and methods for gene trees and species trees.
10:45	Café & croissants
11:15 – 12:45	M.BLUM **-** Approximate Bayesian Computation: algorithms, theory and applications.
12:45	Déjeuner
14:30 – 16:10 (20MIN)	A.LAMBERT - Coalescent point processes and phylogenies. S.GRAVEL - Methods and challenges in the analysis of admixed human genomes. J.SIRÉN - Inference on population trees by approximating wright-fisher diffusions. G.BAELE - Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty L.N.ANDERSEN - Inferring population histories in the IM-model.
19:00 – 20:00	Apéritif & posters(1)
20:00	Dîner

WEDNESDAY 20

09:15 – 10:45	M.STEEL - Probabilistic models of evolutionary trees.
10:45	Café & croissants
11:15 – 12:45	C.ANE **-** Reconstructing species trees, concordance trees and testing the coalescent model.
12:45	Déjeuner
14:00 – 20:00	Balade - Canoë
20:00	Dîner

THURSDAY 21

09:15 – 10:45	E.SUSKO **-** Testing phylogenies.
10:45	Café & croissants
11:15 – 12:45	O.EULENSTEIN - Supertrees and phylogenomics.
12:45	Déjeuner
14:30 – 16:10 (20MIN)	L.ZHANG - Reconciliation of gene and species trees with polytomies. T.WIEHE - Combinatoric and probabilistic properties of coalescent trees and relatives. M.S.BANSAL - Reconciliation revisited : towards faster and more accurate inference of gene family evolution by duplication, transfer, and loss E.BUZBAS - Approximate Bayesian computation when simulating data is difficult. F.ROUSSET - Likelihood-based inference under spatial structure and demographic changes : the Migraine project
19:00 – 20:00	Apéritif & posters(2)
20:00	Dîner & Fête de la musique

FRIDAY 22

09:25 – 10:45 (20MIN)	O.FRANÇOIS - A unified framework for inferring population genetic structure and gene-environment associations. P.PUDLO - Empirical likelihood for Bayesian inference in population genetics. J.BERTL - Approximate inference for high dimensional population genetic models using stochastic gradient methods. E.TANNIER - Evolution of gene neighborhoods within reconciled phylogenies
10:45	Café & croissants
11:15 – 12:45	A.HOBOLTH **-** Understanding the coalescent-with-recombination in time and space: A review and unifying framework.
12:45	Déjeuner
14:00 – 15:30	A.STAMATAKIS - High Performance Phylogenetics.
16:00	The end! Bus to Montpellier

MATHEMATICAL AND COMPUTATIONAL EVOLUTIONARY BIOLOGY

- CONTENTS -

- MONDAY 18 -

INVITED TALKS (1h30)

RECENT PROGRESS ON PHYLOGENETIC NETWORKS.

Vincent Moulton (University of East Anglia, UK)

It is becoming increasingly clear that the evolutionary history of certain organisms (e.g. plants, viruses, bacteria) is not always best represented by a binary leaf-labelled tree. This is due to underlying evolutionary processes such as horizontal gene transfer, recombination and hybridization. Phylogenetic networks provide a framework for exploring and visualizing the complex patterns that can arise from such processes. Even so, they can be complicated structures to construct and understanding how to unravel their complexity for has led to several fascinating problems and results in computer science and mathematics. In this talk we present an overview of phylogenetic networks and discuss some new directions in this rapidly developing area of computational biology.

ABC (APPROXIMATE BAYESIAN COMPUTATION) METHODS TO MAKE INFERENCE ABOUT POPULATION HISTORY FROM MOLECULAR DATA: PRINCIPLES AND APPLICATIONS.

Arnaud Estoup (INRA – CBGP, FR)

One prospect of current biology is that molecular data will help us to reveal the complex demographic histories and processes that have acted on natural populations. The extensive availability of various molecular markers and increased computer power have promoted the development of inferential methods and associated softwares. Among these novel methods, Approximate Bayesian Computation method (ABC) is increasingly used to make inferences from large datasets for complex models in population and evolutionary biology. Briefly, ABC constitutes a recent approach to carrying out model-based inference in a Bayesian setting in which model likelihoods are difficult to calculate (due to the complexity of the models considered) and must be estimated by massive simulations. In ABC, the posterior probabilities of different models and/or the posterior distributions of the demographic parameters under a given model are determined by measuring the similarity between the observed dataset (i.e. the target) and a large number of simulated datasets; all raw datasets (*i.e.* multilocus genotypes or individual sequences) are summarized by so called summary statistics. In this talk I will briefly explain the main principles and advantages of (standard) ABC methods to make inferences in the context of complex evolutionary scenarios. I will then tackle the question of using ABC in practice. To this aim, I will first mention the different steps that need to be carried out to obtain a robust ABC analysis. To illustrate that point, I will take the example of DIYABC, an integrated software for inferring population history with ABC. I will then detail two examples of DIYABC-based analyses on real molecular data sets: (*i*) inferences about the worldwide routes of invasion of the ladybird Harmonia axydis, and (*ii*) inferences about the population history of pygmy populations in Western Africa.

ORAL PRESENTATIONS (20min)

EFFICIENT LYING WITH SIMULATIONS

David Bryant

Simulation experiments are used throughout phylogenetics and bioinformatics to make model comparisons, promote new methods and lobby for particular hypotheses. Simulations can transform subjective design decisions into seemingly objective results. One of the key design decisions is the choice and range of values used for model parameters. We decribe a Markov chain Monte-Carlo (MCMC) framework for conducting simulation experiments. It uses MCMC to sample parameter values in the simulations, thereby improving efficiency and (potentially) transparency. We illustrate the framework with applications to phylogenetics and archaeology.

HOW MUCH DO SPECIES LIMIT EACH OTHER?

Tanja Stadler (1), Gabriel Leventhal (1); Rampal Etienne (2)
(1) ETH Zurich, Switzerland (2) University of Groningen, Netherlands;

In this talk I introduce new methodology which allows testing to which extend speciation and extinction rates are influenced by the abundance of related species (density-dependent speciation and extinction). I use the new likelihood methods to investigate macroevolutionary patterns and importance of competition for a variety of different species clades: mammals, ants and birds. All methods are available within the R package TreePar.

THE EFFECT OF MULTIPLE SEQUENCE ALIGNMENT METHODOLOGY ON DOWNSTREAM EVOLUTIONARY ANALYSES

Ben Blackburne; Simon Whelan
University of Manchester

Many molecular evolutionary analyses start with a multiple sequence alignment, which is usually accepted as known despite wide recognition that errors may impact downstream phylogenetic analysis. Many statistical methods in molecular evolution have been developed to (e.g.) estimate phylogenetic trees or infer adaptation, but these methods are dependent on the accuracy of sequence alignment. Several studies have demonstrated that the results obtained are dependent on the sequence alignment chosen. In cases where systematic error occurs, these differences can be attributed to non-homologous characters being placed together.

To characterize properties of multiple sequence alignment and its effect on downstream evolutionary analysis we examine 200 sets of sequences extracted from The Adaptive Evolution Database, with strict sampling criteria to ensure high quality sequences and reasonable evolutionary divergence. For each set of sequences we apply a range of out-of-the-box popular multiple sequence alignment tools, splitting broadly into 'algorithm-based' aligners (e.g. ClustalW; Muscle; ProbCons; MAFFT; T-Coffee) and phylogenetically-aware aligners (Prank and BAliPhy). We also include samples from the posterior distribution of the statistical aligner BAli-Phy to quantify the degree of uncertainty associated with alignment. We apply recently developed metrics to investigate the similarities and differences between these alignments, finding under multidimensional scaling that algorithm-based and phylogenetically-aware aligners tend to form discrete clusters.
To investigate the effect of alignment on downstream methods we examine two common evolutionary analyses: the inference of a maximum-likelihood tree and a test for adaptive evolution (M7 vs M8 in PAML). For tree estimation algorithm-based and phylogenetically-aware aligners tend to yield noticeably different results, and the distances between alignments seems reasonably correlated with the geodesic distance between trees. To examine the effect of alignment on the inference of adaptive evolution we develop a conservative marginal likelihood approach for integrating across the uncertainty in alignment, based on samples from the statistical aligner BAliPhy. We find that positive results from phylogenetically-aware aligners tend to agree with the results from our marginal likelihood approach, with moderate numbers of additional inferences. Positive results from algorithm-based aligners tend to include those from the marginal likelihood approach, but also include large numbers of additional inferences.

CONSTRUCTING MINIMAL PHYLOGENETIC NETWORKS FROM SOFTWIRED CLUSTERS IS FIXED PARAMETER TRACTABLE

Steven Kelk (1) · Celine Scornavacca (2)
(1) S. Kelk Department of Knowledge Engineering (DKE), Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands.
(2) ISEM, UMR 5554, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier, France
Here we show that, given a set of clusters C on a set of taxa X, where |X| = n, it is possible to determine in time f (k)·poly(n) whether there exists a level-<= k network (i.e. a network where each biconnected component has reticulation number at most k) that represents all the clusters in C in the softwired sense, and if so to construct such a network. This extends a polynomial time result from [1]. By generalizing the concept of 'level-k generator' to general networks, we then extend this fixed parameter tractability result to the problem where k refers not to the level but to the reticulation number of the whole network.
[1] Steven M. Kelk, Celine Scornavacca, Leo van Iersel, "On the Elusiveness of Clusters," IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 517-534, March/April, 2012

PHYLOGENETICALLY DECISIVE TAXON COVERAGE

Mareike Fischer, Ernst-Moritz-Arndt
University Greifswald
In a recent study, Sanderson and Steel defined and characterized phylogenetically decisive sets of taxon sets. A set is called phylogenetically decisive if regardless of the trees chosen for each of its taxon sets, as long as these trees are compatible with one another, their supertree is always unique. It remained unclear whether the decision if a set of taxon sets is phylogenetically decisive can always be made in polynomial time. This question was one of the 'Penny Ante' prize questions of the Annual New Zealand Phylogenetics Meeting 2012. In my talk, I will explain phylogenetic decisiveness and demonstrate a new characterization for it, which then leads to a polynomial time algorithm both for the (simpler) rooted case as well as for the (more complicated) unrooted case.

- TUESDAY 19 -

INVITED TALKS (1h30)

MODELS AND METHODS FOR GENE TREES AND SPECIES TREES.

Noah Rosenberg (University of Michigan, US)
Coalescent models provide a framework for exploring the properties of species tree inference algorithms in the presence of gene tree discordance. A simple coalescent model generates a probability distribution for gene trees evolving along the branches of a species tree; this distribution can then be used to analyze the performance of methods for inferring species trees from gene trees. In this talk, I will investigate the theory of gene tree probabilities and the application of those probabilities in theory- and simulation-based analyses of the properties of species tree inference methods. I will conclude with an example on North American pines that illuminates some of the results obtained theoretically.

APPROXIMATE BAYESIAN COMPUTATION: ALGORITHMS, THEORY AND APPLICATIONS.

Michael Blum (CNRS – TIMC, FR)
Approximate Bayesian computation refers to emerging statistical techniques that do not require likelihood computations for statistical inference. ABC relies on simulations, which makes it particularly suitable for calescent modeling where many simulation tools have been developed. In this talk, I will provide an historical perspective on ABC going from the frequentist premise of ABC to

the more involved regression-based and adaptive methods. Based on a simple example and on theoretical arguments, I will show that regression-based methods can considerably improve the inference obtained with the rejection algorithm. Part of ABC success stems from the possibility to compare the statistical support of different models. Such model selection tools have been severely criticized and I will describe these criticisms and how to address them. In the last part of my talk, I will describe three recent applications of ABC. These applications will be the occasion to stress the importance of model checking, a (sometimes) neglected aspect of Bayesian data analysis.

ORAL PRESENTATIONS (20min)

COALESCENT POINT PROCESSES AND PHYLOGENIES

Amaury Lambert, Tanja Stadler
(1) UPMC Univ Paris 06 - Collège de France (2) ETH Zürich

A coalescent point process is a planar coalescent tree where the coalescence times between two consecutive leaves are independent, identically distributed random variables.

We consider all branching phylogenetic tree models where: 1) lifetimes are independent, they are not necessarily exponentially distributed and their distribution can possibly be time-inhomogeneous; 2) the birth rate possibly also depends on the time variable; 3) species sampling can be incomplete; and we show that under any such models, the tree spanned by (sampled) extant species is always a coalescent point process.

Then we characterize the common distribution of coalescence times in the following three cases: 1) lifetime distributions and birth rate do not depend on time; 2) the time-inhomogeneity of lifetime distributions is due to inhomogeneous, instantaneous death rates; 3) the time inhomogeneity of lifetime distributions is due to a series of bottlenecks.

METHODS AND CHALLENGES IN THE ANALYSIS OF ADMIXED HUMAN GENOMES.

Simon Gravel; Jeffrey M. Kidd; Jake K. Byrnes; Carlos D. Bustamante
M-327-Genetics, 259 Campus Drive-ALWAY BLDG

A substantial proportion of humans are "admixed", in the sense that their recent ancestors belong to statistically distinct groups. This needs to be accounted for if unbiased inference and associations are to be performed. We present a diversity of methods for the analysis of whole-genome sequence data from admixed individuals, and apply them to 50 genomes sequenced by Complete Genomics, including 4 Mexican-Americans, 4 African-Americans and 2 individuals from Puerto Rico, together with SNP genotype data from hundreds of additional samples.

Many methods have been presented recently to infer the population of origin of specific loci along the genomes of admixed individuals, leading to inferred mosaics of ancestry. We first propose a simple Markov model that relates the time-dependent migration history to the inferred patterns of local ancestry. We use this framework to infer the timing of admixture and to differentiate between punctual and continuous models of migration: using demographic models that are consistent with both historical records and genetic data, we find evidence for continuous migration patterns in both Mexican and African-American populations.

We also propose models to study the longer-term evolution of the ancestral populations, by considering the allele frequency distribution, pairwise TMRCA's, and by a simple extension of the recently introduced Pairwise Sequencial Markov Chain approach for demographic inference. The inferred source population demographic histories are in broad agreement with previous results for European and West-African populations, and the inferred demography for the Native source population closely follows the European one until about 20,000 years ago. Taken together, whole genome sequencing and local ancestry assignment therefore permit inferences about long-term histories of unsampled ancestral populations and highlights recent historical demographic processes that altered patterns of variation observed in admixed populations.

INFERENCE ON POPULATION TREES BY APPROXIMATING WRIGHT-FISHER DIFFUSIONS.

Jukka Sirén
University of Helsinki

We consider the problem of inferring population / species trees from a wide variety of genetic markers. It has received considerable attention in the last two decades as it has been realized that gene trees obtained from phylogenetic analyses of molecular data might not adequately reflect the evolutionary history of the species. We concentrate here on methods which avoid the need to compute gene tree probabilities explicitly. The history of such methods can be traced to the early developments by Cavalli-Sforza, Edwards, Felsenstein and others in the 1960's using diffusion approximations to Wright-Fisher models.

We have developed approximations to the Wright-Fisher diffusions for different types of genetic markers including SNPs and MLST sequences. We compute properties of the discrete Wright-Fisher models before transforming to the diffusion scale. The diffusions are approximated using Dirichlet distributions and their generalizations, making the methods computationally manageable and suitable for large data sets.

Recently, a coalescent based approach was introduced for species tree inference from biallelic loci with mutation. We derive a comparable method by approximating the corresponding Wright-Fisher diffusion. By considering directly species level properties our model facilitates inferences with data sets harboring large numbers of individuals. We present comparison between the coalescent and diffusion based approaches with simulated and real data sets.

D Bryant, R Bouckaert, J Felsenstein, NA Rosenberg, A RoyChoudhury. 2012 Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Molecular Biology and Evolution (in press).
J Siren, P Marttinen, J Corander. 2011. Reconstructing Population Histories from Single Nucleotide Polymorphism Data. Molecular Biology and Evolution. 28:673--683.
J Siren, WP Hanage, J Corander. 2012. Inference on Population Histories by Approximating Infinite Alleles Diffusion. Under revision.

IMPROVING THE ACCURACY OF DEMOGRAPHIC AND MOLECULAR CLOCK MODEL COMPARISON WHILE ACCOMMODATING PHYLOGENETIC UNCERTAINTY

Guy Baele (1), Philippe Lemey (2), Trevor Bedford (3), Andrew Rambaut (4), Marc Suchard (5) and Alexander Alekseyenko (6) Rega Institute; Kapucijnenvoer 33 blok i - box 7001; 3000 Leuven; Belgium

Recent developments in marginal likelihood estimation for model selection in the field of Bayesian phylogenetics and molecular evolution have emphasized the poor performance of the harmonic mean estimator (HME). Although these studies have shown the merits of new approaches applied to standard normally distributed examples and small real-world data sets, not much is currently known concerning the performance and computational issues of these methods when fitting complex evolutionary and population genetic models to empirical real-world data sets. Further, these approaches have not yet seen widespread application in the field, due to the lack of implementations of these computationally demanding techniques in commonly-used phylogenetic packages. We have investigated the performance of some of these new marginal likelihood estimators, specifically, path sampling and stepping-stone sampling for comparing models of demographic change and relaxed molecular clocks, using synthetic data and real-world examples for which unexpected inferences were made using the HME. Given the drastically increased computational demands of path sampling and stepping-stone sampling, we also investigate a posterior simulation-based analogue of Akaike's information criterion (AICM) through Markov chain Monte Carlo (MCMC), a model comparison approach which shares with the HME the appealing feature of having a low computational overhead over the original MCMC analysis. We confirm that the HME systematically overestimates the marginal likelihood and fails to yield reliable model classification and show that the AICM performs better and may be a useful initial evaluation of model choice but that it is also, to a lesser degree, unreliable. We show that path sampling and stepping-stone sampling substantially outperform these estimators and adjust the conclusions made concerning previous analyses for the three real-world data sets that we reanalyzed. The methods used in this paper are now available in BEAST, a powerful user-friendly software package to perform Bayesian evolutionary analyses.

INFERRING POPULATION HISTORIES IN THE IM-MODEL.

Lars Nørvang Andersen; Thomas Mailund; Asger Hobolth.
Bioinformatics Research Centre Aarhus University C.F. Møllers Allé 8 DK-8000 Aarhus C Denmark

We present a framework for estimating parameters in an Isolation-with-migration-model, intended for inference of parameters based on whole-genome alignment data. The involved parameters are the split time, migration rates between demes, and effective population sizes, and our method allows for joint estimation of these parameters.
We derive explicit analytical results and use these to perform maximum likelihood estimation, under the assumption of free recombination. We show the magnitude of the inherent state-space explosion and demonstrate how to deal with it. Furthermore, we demonstrate how to use matrix exponentials for efficient computation.
Finally, we consider both an infinite-sites model and a general Markov model for nucleotide substitutions as mutation models, where the general Markov model is used for site pattern analysis. We compare the performance of our method for the two types of mutation models in different scenarios involving the underlying true parameters, and in different sub-models, and asses when we are able to recover the true parameters, and when this is difficult due to a flat log-likelihood curve.
On computing the coalescence time density in an isolation-with-migration model with few samples,Genetics, 187, 1241--1243.
Nielsen, R. and J. Wakeley (2001)
Distinguishing migration from isolation: A markov chain monte carlo approach, Genetics, 158, 885--896.
Wang, Y. and J. Hey (2010)
Estimating divergence parameters with small samples from a large number of loci, Genetics, 184, 363—379

- Wednesday 20 -

INVITED TALKS (1h30)

Probabilistic models of evolutionary trees.

Mike Steel (1) Arne Mooers (2)
(1) University of Canterbury (2) Simon Fraser University

The 'shape' of a reconstructed evolutionary tree depends, in part, on the underlying random process of speciation, extinction, and taxon sampling. In this talk, I describe some results (both classical and recent) concerning the distribution of tree shapes that arise under various neutral macro-evolutionary models. I also outline the implications of these results for phylogenetic inference.

Reconstructing species trees, concordance trees and testing the coalescent model

Cécile Ané (University of Wisconsin, US)

Conflict among gene trees is not new in systematics, but can be quantified with increasing accuracy with the growth of sequencing power. In the first part, I will review the various statistical methods that have been developed to handle discordant gene trees. I will highlight the methods' assumptions, advantages and disadvantages that have been uncovered in a number of recent simulation studies. Most gene tree / species tree methods assume that the discordance among gene trees is due to incomplete lineage sorting, as mathematically explained by the multi-species coalescent model from population genetics. The coalescent model will be presented with some of its most important properties for species tree reconstruction.
In the second part, I will consider the goal of reconstructing not only the species tree but also the "phylome": the whole distribution of gene trees along the genome. The phylome contains information about the main signal of vertical inheritance (the species tree) as well as the horizontal signal and the processes that caused gene tree conflict. Bayesian concordance analysis can be used to quantify the genetic support for vertical as well as "horizontal" relationships, and to test the null hypothesis that the multispecies coalescent model is sufficient to explain the observed conflict among gene trees. The link between species trees and coalescent trees will be clarified in light of the "too-greedy" zone of the consensus method.

- Thursday 21 -

INVITED TALKS (1h30)

Testing phylogenies.

Edward Susko (Dalhousie University, CA)

Topologies are unusual parameters that might be considered as discrete or, when edge-lengths are taken into account, as non-convex regions of a continuous parameter space. Not surprisingly, standard statistical methods for testing and interval estimation are not easily applied to tree inference.
In the first pat of the talk we will formulate the problem of testing or, equivalently, confidence region construction. We review some of the currently available methods with examples and discuss issues of appropriate hypotheses, multiplicity of tests and selection bias.
While not every phylogenetic analysis includes tests of topology, almost all include some measure of support for the splits present in estimated or hypothesized trees of interest. Bootstrap support or bootstrap probability (BP) is by far the most frequently used measure. We discuss issues of interpretation for BP and present adjusted versions with more conventional interpretations. Approaches to determining support for branches based on likelihood ratio testing are presented as worthwhile alternatives. The second most frequently used measure of support for splits are posterior probabilities. We conclude this portion of the talk with a brief discussion of their properties
We conclude with some notes about speeding up bootstrapping and some additional observations about the properties of methods.

Supertrees and phylogenomics.

Olivier Eulenstein (Iowa State University, US)

A supertree is a phylogenetic tree that summarizes a collection of input trees that share some but not necessarily all of their terminal taxa. The interest in supertrees is largely due to their potential use for constructing large phylogenies of major clades of the tree of life and synthesizing phylogenetic estimates derived from disparate types of data.
In this presentation I will motivate supertrees, describe some of the major supertree problems, and algorithms or heuristics to address these problems. I also outline a classification of supertree problems and concepts to enhance the effectiveness supertree problems.
In the last part, I will address challenges brought up by polyploid species. First, the origins of allopolyploid species introduce reticulation events which are typically not handled by specie tree methods. Second, polyploid species have several alleles at each

locus, but it is unknown which alleles should be grouped together as coming from the same parental origin, particularly in the presence of gene tree conflict. I will present an approach to harness gene tree / species tree methods to clarify the evolution of polyploids.

ORAL PRESENTATIONS (20min)

RECONCILIATION OF GENE AND SPECIES TREES WITH POLYTOMIES

Louxin Zhang
National University of Singapore

Millions of genes in the modern species belong to only thousands of `gene families'. A gene family includes instances of the same gene in different species and duplicate genes in the same species. Genes are gained and lost during evolution. With advances in sequencing technology, researchers are able to investigate the important roles of gene duplications and losses in adaptive evolution. Because of gene complex evolution, ortholog identification is a basic but difficult task in comparative genomics. A key method for the task is to use an explicit model of the evolutionary history of the genes being studied, called the gene (family) tree. It compares the gene tree with the evolutionary history of the species in which the genes reside, called the species tree, using the procedure known as tree reconciliation. Reconciling binary gene and specific trees is simple. However, both gene and species trees may be non-binary in practice and thus tree reconciliation presents challenging problems. Here, non-binary gene and species tree reconciliation is studied in a binary refinement model.

The problem of reconciling arbitrary gene and species trees is proved NP-hard even for the duplication cost. We then present the first efficient method for reconciling a non-binary gene tree and a non-binary species tree. It attempts to find binary refinements of the given gene and species trees that minimize reconciliation cost. Our algorithms have been implemented into a software to support quick automated analysis of large data sets.

COMBINATORIC AND PROBABILISTIC PROPERTIES OF COALESCENT TREES AND RELATIVES

Thomas Wiehe, Filippo Disanto
Institut fuer Genetik, Zuelpicher Strasse 47a, 50674 Koeln, Germany

Evolutionary trees can be viewed as a special class of rooted binary trees. Depending on the choice of an appropriate equivalence relation, coalescent trees, shape trees, labelled coalescent trees or phylogenetic trees emerge. We present results on the enumeration and the probability distribution of different classes of trees and their relationship to certain classes of alternating down-up permutations. Some results on tree shape properties are obtained with the help of generating functions. Finally, we will talk about tree distance between a pair of coalescent trees and its relationship to recombinational distance along a chromosome.

RECONCILIATION REVISITED: TOWARDS FASTER AND MORE ACCURATE INFERENCE OF GENE FAMILY EVOLUTION BY DUPLICATION, TRANSFER, AND LOSS

Mukul S. Bansal (1); Eric J. Alm (2); Manolis Kellis (1,3)
(1) Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA; (2) Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, USA; (3) Broad Institute of MIT and Harvard, Cambridge, USA.

Gene family evolution is driven by evolutionary events like speciation, gene duplication, horizontal gene transfer, and gene loss, and inferring these events in the evolutionary history of a given gene family is a fundamental problem in comparative and evolutionary genomics with numerous important applications. This problem is typically solved using a reconciliation framework, where the input consists of a gene family phylogeny and the corresponding species phylogeny, and the goal is to parsimoniously reconcile the two by postulating speciation, gene duplication, horizontal gene transfer, and gene loss events. This reconciliation problem is referred to as Duplication-Transfer-Loss (DTL) reconciliation and has been extensively studied in the literature. Yet, the biological utility of this powerful approach remains rather limited and several important issues remain to be addressed to allow it to be easily and usefully applied to large-scale biological datasets. For instance, despite tremendous advances in the algorithmics of the problem, even the fastest existing algorithms for DTL-reconciliation are too slow to reconcile large gene families and for use in more sophisticated biological applications such as gene tree or species tree reconstruction. Similarly, existing solutions do not properly address the fact that there are often multiple optimal reconciliations, or that gene trees often contain errors.

In this talk we present recent results that address these issues and improve upon the current state-of-the-art for DTL-reconciliation in several important ways. These include (i) efficient algorithms for the DTL-reconciliation problem that are dramatically faster than existing algorithms, both asymptotically and in practice, (ii) consideration of distance-dependent transfer costs that allow for a more accurate reconciliation, (iii) dealing cleanly with multiple optimal solutions by sampling the space of all optimal reconciliations uniformly at random and aggregating the results, (iv) methods for handling errors in gene tree topologies, and other enhancements. These improvements greatly enhance the speed and accuracy of DTL-reconciliation, making it possible to use DTL-reconciliation for performing rigorous and accurate evolutionary analyses of even large gene families, and enabling its use in sophisticated applications like reconciliation-based gene and species tree reconstruction.

APPROXIMATE BAYESIAN COMPUTATION WHEN SIMULATING DATA IS DIFFICULT

Erkan Buzbas, Noah Rosenberg
Department of Biology Stanford University 371 Serra Mall Stanford, CA 94305-5020 USA

Approximate Bayesian computation (ABC) methods perform approximate inference on the posterior distribution of parameters without explicitly evaluating the model likelihoods. Central to the success of ABC methods is inexpensive simulation of data sets under the model of interest, which is not always feasible when the stochastic model generating the data is complex. A common practice therefore is to model the system of interest at a level of complexity that statistical inference is computationally feasible. This approach, however, often implied using a less realistic mechanistic model than originally conceived by a researcher. We present an extension of ABC that performs inference without necessarily sacrificing model complexity at a structural level, at the cost of a less precise inference under the original model complexity. Our method exploits information provided by a limited number of simulated data sets generated under the complex model. The novelty that we bring to the ABC framework is to use a nonparametric model for the data, in which the data sets simulated under the complex model are treated as background information. This approach allows us to simulate approximate data sets that are otherwise unavailable. We show that for appropriately chosen nonparametric models and priors, the posterior distribution targeted by our method converges to the posterior distribution sampled by standard ABC methods as the number of simulated data sets and the sample size of the observed data set increase. We demonstrate the method to perform inference on the parameters of a model from genetics and a model describing population demography.

LIKELIHOOD-BASED INFERENCE UNDER SPATIAL STRUCTURE AND DEMOGRAPHIC CHANGES: THE MIGRAINE PROJECT

Francois Rousset(1) Raphaël Leblois(2) Beeravolu, C.R.(1) et Pudlo P.(3)
(1) Institut des Sciences de l'Evolution CNRS UMR 5554, Université Montpellier 2; (2) Centre de Biologie et Gestion des Populations Naturelles INRA, Campus International de Baillarguet, Montferrier-sur-Lez, France (3) I3M, Université de Montpellier II - CC 051 34095 MONTPELLIER Cedex 05 ;

We consider the likelihood-based inference of demographic and mutation parameters under (i) scenarios of population structure and (ii) of change in population size, based on allelic type information such as microsatellite genotypes. de Iorio and Griffiths (2004) have defined a class of importance sampling algorithms well-suited for the first scenario. We will present the current state of the 'Migraine' project where we use such algorithms to implement, validate and investigate the robustness of likelihood-based inferences under both demographic scenarios.

References and link:

Migraine software home page: http://kimura.univ-montp2.fr/~rousset/Migraine.htm
De Iorio, M. and Griffiths, R. C. (2004). Importance sampling on coalescent histories. I and II. Adv. Appl. Prob. 36, 417-454.
De Iorio, M., Griffiths, R.C., Leblois, R., Rousset, F. (2005) Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. Theor. Pop. Biol. 68: 41-53.
Rousset, F., Leblois, R. (2007) Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model mis-specification. Mol. Biol. Evol. 24: 2730-2745.
Rousset, F., Leblois R. (2012) Likelihood-based inferences under isolation by distance: two-dimensional habitats and confidence intervals. Mol. Biol. Evol. 29: 957-973. dx.doi.org/10.1093/molbev/msr262

- FRIDAY 22 -

INVITED TALKS (1h30)

UNDERSTANDING THE COALESCENT-WITH-RECOMBINATION IN TIME AND SPACE: A REVIEW AND UNIFYING FRAMEWORK

Asger Hobolth (Aarhus University, DK)

The coalescent-with-recombination is a fundamental process for understanding genetic variation, and simulation strategies and inference techniques for the process continue to improve. Simulation procedures are at a particularly high level: large genetic data sets can be simulated for many complex demographic scenarios. Methodologies for genome-wide statistical inference in the coalescent-with-recombination process have been developed in the past few years. The methods are often based on a hidden Markov model along the sequence alignment, where the hidden states are trees and the emissions are alignment columns. It remains a challenge to model the transitions between trees. The main aim of this talk is to review and compare different versions and approximations of the coalescent-with-recombination. In particular we describe in detail the process in both time and space, discuss the assumptions involved for the coalescent-with-recombination to be a spatial Markov chain in tree space, and solve the issues that arise when space is discretized.

HIGH PERFORMANCE PHYLOGENETICS.

Alexandros Stamatakis (Heidelbelg, DE)

Computing the likelihood function on trees represents the key computational/technical challenge for designing fast Maximum Likelihood and Bayesian Inference Programs that typically spend more than 90% of their total run time in this function.

The calculation of the phylogenetic likelihood function is both memory- as well as compute-intensive.

Initially, (part 1) I will briefly review how the likelihood is computed on trees and derive a formula for calculating the main memory (RAM) requirements for calculating the likelihood on an alignment with n taxa and m sites/columns.

In part 2, I will discuss generally applicable techniques for optimizing and parallelizing likelihood computations on a single tree on modern parallel computer architectures. I will address common parallel performance problems that are associated to load imbalance among processors and communication bottlenecks. These problems are particularly pronounced on large partitioned multi-gene datasets.

Finally, in Part 3, I will present some of our recent research to alleviate load imbalance problems and discuss three generally applicable techniques for reducing the memory footprints of likelihood calculations.

I will also briefly address current work on optimizing and parallelizing population genetics codes for computing the omega statistic and conducting forward-simulations.

ORAL PRESENTATIONS (20min)

A UNIFIED FRAMEWORK FOR INFERRING POPULATION GENETIC STRUCTURE AND GENE-ENVIRONMENT ASSOCIATIONS

Olivier François, Eric Frichot, Sean Schoville, Guillaume Bouchard

Université Joseph Fourier, TIMC-IMAG UMR 5525, 38402 Grenoble France

Local adaptation through natural selection plays a central role in shaping the genetic variation of populations. A way to investigate signatures of local adaptation, especially when beneficial alleles have weak phenotypic effects, is to identify polymorphisms that exhibit high correlation with environmental variables. However the geographical basis of both environmental and genetic variation can confound interpretation of these associations, as they can also result from genetic drift at neutral loci.

Here we propose an integrated framework based on spatial statistics, population genetics and ecological modeling for scans for signatures of local adaptation from genomic data. We present a novel class of algorithms to detect correlations between environmental and genetic variation that take account background levels of population structure and spatial autocorrelation in allele frequencies generated by isolation-by-distance mechanisms.

Our framework uses probabilistic matrix factorization, a hierarchical Bayesian mixed model in which environmental variables are fixed effects and population structure is introduced as random effects. We implement fast algorithms that simultaneously estimate the scores and loadings, the effects of environmental variables and the local autocorrelation scale parameter. We show that our algorithm can used to 1) correct for spatial autocorrelation when running principal component analysis, and 2) control random effects due to population history when estimating gene-environment correlations. We give examples of applications to simulated data and to human genetic data and climatic variables.

EMPIRICAL LIKELIHOOD FOR BAYESIAN INFERENCE IN POPULATION GENETICS

P. Pudlo (2 & 3); C.P. Robert (1); R. Leblois (2)

(1) University Paris Dauphine & IUF; (2) INRA CBGP; (3) University Montpellier 2

In population genetics, the computation of the likelihood is often a hard problem. Approximate Bayesian computation (ABC) is certainly the most used algorithm to bypass this computation in a Bayesian paradigm (see Beaumont, 2010, for a recent survey). But ABC is quite time consuming and needs massive paralellization to be efficient. In this talk we will present a promising alternative using the empirical likelihood (Owen, 1988).

That last method profiles the likelihood in a nonparametric way using an estimating equation on the unknown parameters. Our proposal relies on the score functions given by the pairwise composite likelihood (Lindsay, 1988) which can be explicitly computed in a large variety of evolutionary scenarii when considering microsatellite loci with the stepwise mutation model (Ohta and Kimura, 1973). Numerical simulations will exhibit that the posterior estimated with our proposal is comparable to the ABC posterior, but that the computation is about thirty times faster.

Beaumont, M. (2010) Approximate Bayesian computation in evolution and ecology. Annu. Rev. Ecol. Evol. Syst. 41: 379-406

Lindsay, B.G. (1988). Composite Likelihood Methods. Contemporary Mathematics 80: 221-239

Ohta, T. and Kimura., M. (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. 22: 201-204.

Owen, A. B. (1988) Empirical Likelihood ratio confidence intervals for a single functional. Biometrika 75: 237-249.

Approximate Inference for High Dimensional Population Genetic Models using Stochastic Gradient Methods

Bertl Johanna, Futschik Andreas
Dept. of Statistics, University of Vienna, Universitaetsstrasse 5/9, 1010 Vienna, Austria

As exact likelihoods are usually hard to obtain in population genetics, methods of approximate inference such as ABC became popular in recent years. In their most basic form, these methods often involve sampling from the parameter space and keeping those parameters that produce data that fit best to the actually observed data. Exploring the whole parameter space however, does not make this approach very efficient in high dimensional problems. Therefore alternative methods have been proposed that aim for exploring the parameter space more efficiently. (Think for instance of approaches that are based on MCMC or importance sampling.)
Here we propose an approach that is based on stochastic gradient methods. By moving along an estimated gradient (or alternatively ascent direction) of the likelihood, the algorithm produces a sequence of estimates that will eventually converge to the maximum likelihood estimate. Analogously it is possible to construct a sequence of estimates converging to the maximum posterior. A classical method in this context is the Kiefer-Wolfowitz algorithm. Besides some theoretical properties, we also investigate the practical performance of our proposed approach. Since it turns out that a good choice of tuning parameters is essential for the performance of the algorithm, we propose some good tuning strategies. We explore the performance of the algorithm and provide a comparison with classical ABC. We will also explore up to which problem dimension our proposed algorithm produces reliable results.

Evolution of gene neighborhoods within reconciled phylogenies

Eric Tannier (1), Sèverine Bérard (1); Bastien Boussau (2,3); Vincent Daubin (2); Gergerly Szollosi (2)
(1) Cirad, Univ Montpellier; (2) CNRS, LBBE, Univ Lyon; (3) Univ California Berkeley

We construct "adjacency phylogenetic trees", which describe the evolution of a neighborhood relation between two genes, by speciation, duplication of one or the two genes, loss of one or the two genes, rearrangement.
We give an algorithm which, given a species tree and a set of gene trees where the leaves are connected by adjacencies, computes an adjacency forest which minimizes the number of gains and losses of adjacencies, and runs in polynomial time.
We use this algorithm to reconstruct contiguous regions of mammalian or plant ancestral genomes. We are thus able to detect duplications involving several genes, and compare the different modes of evolution

- POSTERS (1) -

SYSTEMATIC ERROR IN SEED PLANT PHYLOGENOMICS

Bojian Zhong (1), Oliver Deusch (1), Vadim V. Goremykin (2), David Penny (1) and Peter Lockhart (1)
(1) Institute of Molecular Biosciences, Massey University, NZ (2) IASMA Research Center, Italy

Resolving the closest relatives of Gnetales has been an enigmatic problem in seed plant phylogeny. The problem is known to be difficult because of the extent of divergence between this diverse group of gymnosperms and their closest phylogenetic relatives. Here we investigate the evolutionary properties of conifer chloroplast DNA sequences. To improve taxon sampling of Cupressophyta (non-Pinaceae conifers) we report sequences from three new chloroplast (cp) genomes of Southern Hemisphere conifers. We have applied a site pattern sorting criterion to study compositional heterogeneity, heterotachy and the fit of conifer chloroplast genome sequences to a GTR + G substitution model. We show that non-time reversible properties of aligned sequence positions in the chloroplast genomes of Gnetales mislead phylogenetic reconstruction of these seed plants. When 2250-3000 of the most varied sites in our concatenated alignment are excluded, phylogenetic analyses favour a close evolutionary relationship between the Gnetales and Pinaceae – the Gnepine hypothesis. Our analytical protocol provides a useful approach for evaluating the robustness of phylogenomic inferences. Our findings highlight the importance of goodness of fit between substitution model and data for understanding seed plant phylogeny.

HYBRIDIZATION NETWORKS FOR MULTIPLE TREES

Leo van Iersel and Simone Linz
Science Park 123, 1098 XG Amsterdam, Netherlands; Sand 14, 72076 Tübingen, Germany

It has recently been shown that the NP-hard problem of calculating the minimum number of hybridization events that is needed to explain a set of rooted binary phylogenetic trees by means of a hybridization network is fixed-parameter tractable if an instance of the problem consists of precisely two such trees. We now show that this problem remains fixed-parameter tractable for an arbitrarily large set of rooted binary phylogenetic trees. In particular, we present a quadratic kernel.

VIRUS (OR HOST) GENOTYPE WITH CLINICAL TRAITS IN HIV OR HCV INFECTIONS

Samuel Alizon (1); Christophe Fraser (2); George Shirreff (2,3); Tanja Stadler (3); Jacques Fellay (4); Amalio Telenti (5); Huldrych Günthard (6); Sebastian Bonhoeffer (3)
(1) Laboratoire MIVEGEC (UMR CNRS 5290, IRD 224, UM1, UM2) Montpellier, France;
(2) School of Public Health, Imperial College, London;
(3) Institut f. Integrative Biologie, ETH Zürich;
(4) EPFL Lausanne; (5) Institute of Microbiology, Université de Lausanne;
(6) University Hospital, Zürich

Human viruses such as HIV or HCV (hepatitis C virus) are a major public health concern largely because they evolve rapidly. An open question is to determine whether the variance in symptoms that we observe between patients is due to host or virus genetic factors. Estimating virus control over infection traits is a difficult task because we often ignore the transmission chain (i.e. who infected whom). I will first show that classical phylogenetic comparative approaches can be applied to virus phylogenies of HIV in order to estimate virus control over an infection trait (set-point virus load) [Alizon et al. 2010]. I will also compare the appropriateness of various measures of phylogenetic signal to estimate this virus control [Shirreff, Alizon and Fraser, in prep]. I will then discuss two extensions of this framework. One consists in applying this approach to a categorical trait (whether an infection by HCV is chronic or acute). The other consists in building human genealogies in order to estimate the control of the human genotype over the infection trait.

ORIGINS AND EVOLUTION OF THE ETRUSCANS' DNA

Ghirotto Silvia (1), Tassi Francesca (1), Fumagalli Erica (1,2), Colonna Vincenza (1,3), Lari Martina (4), Rizzi Ermanno (5), Caramelli David (4), Barbujani Guido (1)
(1) Department of Biology and Evolution, University of Ferrara, via Borsari 46, 44100 Ferrara, Italy ;
(2) Department of Ecology and Evolution, UNIL-Sorge, BB 2105, University of Lausanne, CH-1015, Lausanne, Switzerland;
(3) Istituto di Genetica e Biofisica "Adriano Buzzati-Traverso", National Research Council (CNR), Via Pietro Castellino 111, 80131 Napoli, Italy;
(4) Department of Evolutionary Biology, University of Firenze, via del Proconsolo 12, 50122 Firenze, Italy;
(5) Institute for Biomedical Technologies (ITB), National Research Council (CNR), Via F.lli Cervi 93, 20090 Segrate, Milan, Italy

The Etruscan culture is documented in Etruria, Central Italy, from the 7th to the 1st century BC. For more than 2,000 years there has been disagreement on the Etruscans' biological origins, whether local or in Anatolia. Genetic affinities with both Tuscan and Anatolian populations have been reported, but so far all attempts have failed to fit the Etruscans' and modern populations in the

same genealogy. We extracted and typed mitochondrial DNA of 14 individuals buried in two Etruscan necropoleis, analyzing them along with other Etruscan and Medieval samples, and 4,910 contemporary individuals. Comparing ancient and modern diversity with the results of millions of computer simulations, we show that the Etruscans can be considered ancestral, with a high degree of confidence, to the modern inhabitants of two communities, Casentino and Volterra, but not to most contemporary populations dwelling in the former Etruscan homeland. We also show that the genetic links between Tuscany and Anatolia date back to at least 5,000 years ago, strongly suggesting that the Etruscan culture developed locally, without a significant contribution of recent Anatolian immigrants.

DAWG 2.0: NEW METHODS FOR SEQUENCE SIMULATION.

Reed A. Cartwright
Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, USA

Simulations of sequence evolution are an important component of research in molecular evolution, comparative genomics, and bioinformatics. Existing simulation technologies cannot reliably replicate critical characteristics of real sequences. Dawg 2.0 will provide a framework to solve these problems. Currently the gold standard of simulating sequence evolution is using the Gillespie algorithm to process both point mutations and indels as a sequence evolves along a branch. Dawg 2.0 turns the Gillespie algorithm on its side and processes events as they happend along a sequence. Thus the new algorithm avoids complex bookkeeping and search trees, effeciently processing both indels and rate variation. Speed-ups of 1200% are seen for some tasks. Significant new features of Dawg 2.0 include support for nucloetide, protein, and codon evolutionary models, and model heterogenity along a tree and sequences.

DEMOGRAPHIC INFERENCE USING SKYLINE PLOTS ON APPROXIMATE BAYESIAN COMPUTATION

Miguel Navascués1; Concetta Burgarella2
1 INRA, UMR CBGP, Campus international de Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez cedex, France; 2 INRA, UMR AGAP, Montpellier, France

Bayesian Skyline Plots (BSPs) are representations of the posterior probability density of the effective population size in function of time; i.e. a graphical representation of the fluctuation of the effective population size with time based on the estimates obtained from Bayesian inference. The interest of BSPs is that they allow to infer gradual changes of the effective population size without the need of a specific mathematical function determining the shape of the demographic change. For instance, a population expansion can be well characterized by a BSP either if the change had occurred instantaneously, exponentially or logistically. Currently, the only implementation of this analysis has been done within the MCMC-based estimation of likelihood approach and is restricted to non-recombining DNA sequence data. We have explored how to implement BSP within the approximate Bayesian computation (ABC) framework, with promising preliminary results. An implementation in ABC allows to obtain BSP from any multilocus molecular data, e.g. recombining DNA sequence data, microsatellites, AFLPs or SNPs.

COMPUTING THE JOINT DISTRIBUTION OF TREE SHAPES AND TREE DISTANCES FOR MULTIPLE TREE INFERENCE WITH TREE METRIC BASED MODELS

(1). Yujin Chung; (2) Cécile Ané
(1)-(2) Department of Statistics (2) Department of Botany, Unversity of Wisconsin - Madison, USA

Recombination events and other biological processes can cause the topologies of phylogenetic trees to be discordant for different genes. The dissimilarity among gene trees can be incorporated in a model to improve the accuracy of tree inference, when we seek to simultaneously detect recombination breakpoints along an alignment and infer phylogenetic trees of segments defined by the recombination breakpoints. A Gibbs distribution can be used to describe the dissimilarities among gene trees in terms of the Robinson-Foulds (RF) distances between neighboring gene trees. Modeling the RF distance between tree topologies of neighboring segments allows the detection of recombination breakpoints between short segments with similar tree topologies. When taking into account the RF distance between trees of neighboring segments, a major difficulty is the calculation of the "partition function", which works as a normalizing constant for the Gibbs distribution on trees. When the partition function is overlooked or miscalculated, an incorrect maximum likelihood estimate or an incorrect Bayesian posterior distribution may be obtained. Calculating the partition function in the naive way is computationally prohibitive.

We derive here an algorithm to calculate the partition function exactly, based on the calculation of the joint distribution of the tree shapes and the RF distance between two random trees. We derive this joint distribution through a system of generating functions, which is similar to the algorithm proposed by Bryant and Steel (2009). We also propose approximations to the partition function, which are computationally fast and accurate. Finally, we tie this work back to the problem of recombination breakpoint detection and tree reconstruction and its benefits to the development of correct MCMC Bayesian estimation.

Reference:

David Bryant and Mike Steel. 2009. Computing the Distribution of a Tree Metric. IEEE/ACM Trans. Comput. Biol. Bioinformatics 6, 3, 420-426.

UNDERSTANDING THE EFFECT OF REFERENCE TREE IN THE PHYLOGENETIC PLACEMENT OF METAGENOMIC DATA

Eliza Loza, Nick Goldman
EMBL-European Bioinformatics Institute, Hinxton Cambridge, UK

Metagenomics is the study of microbial DNA directly extracted from a habitat (e.g. agricultural soil, ocean water, or the human gut). When using second-generation sequencing technologies, the observed data are thousands of short DNA sequences that originate from the genomes of the microorganisms that populate the sampled habitat. A typical objective in a metagenomic study is to associate the metagenomic fragments with the operational taxonomic units (e.g. species, strains, or populations) or functions of their origin. Some existing methods build upon the strength and reliability of likelihood-based phylogenetic approaches, such as the placement of metagenomic sequences onto a reference phylogeny (e.g. [1,2]).

The input in phylogenetic placement consists of a reference tree, a set of one or more marker genes, and the metagenomic data. The marker genes are used to identify 'phylogenetically reliable' sequences within the totality of the sample using, for example, HMMER searches. The parameter of primary inferential interest is the set of assignments of the metagenomic fragments onto the reference tree. Once the assignments are observed, their distribution along the edges of the reference tree is used as an indication of biodiversity and, in some cases, of relative abundance of the microbes in the habitat.

A variety of approaches has been used to arrive at the reference trees used in applications of phylogenetic placement, including a reference tree of all the organisms whose genome has been fully sequenced, and trees constructed from alignments of concatenated marker genes. However, no study that we are aware of has assessed the effects of the reference tree on the inferential process. In studies of exceptionally rich and biodiverse habitats assignments of metagenomic fragments onto a reference tree will represent habitat biodiversity as poorly/adequately as the reference tree itself represents life biodiversity.

In the first stage of our study, we have used computer simulation to quantify phylogenetic-placement accuracy given a choice of reference tree. Phylogenetic placement was conducted using the pplacer software package [1], and three different measures of accuracy were used. The output of this stage is a profile of placement accuracy on the selected reference tree. Once this accuracy profile can be derived, one can vary the reference tree to study the effects of these variations on placement accuracy.

Our method aims to provide guidelines on the reference tree to use in a metagenomic study. If, for instance, improvement in the accuracy profile is due to deeper resolution in a particular clade of the reference tree, the investigator could allocate resources to obtain additional reference sequence data for that clade. We present preliminary results from our method.

[1] Matsen et al. BMC Bioinformatics 2010, 11(1):538.
[2] Stark et al. BMC Genomics 2010, 11:461.
[3] von Mering et al. Science 2007, 315:1126–1130.

MAXIMUM LIKELIHOOD INFERENCE COMBINING SEQUENCE AND ALLELIC MARKERS IN A POPULATION OF VARIABLE SIZE: AN IMPORTANCE SAMPLING APPROACH

Beeravolu, C.R. (1); Leblois, R. (2); Pudlo, P. (3) & Rousset, F (4).
(1)(2)(3) CBGP, Campus International de Baillarguet CS 30016 34988 Montferrier-sur-Lez cedex, France;
(3) I3M, Université de Montpellier II - CC 051 34095 MONTPELLIER Cedex 05 ;
(4) ISEM, Université de Montpellier II - CC 065 34095 MONTPELLIER Cedex 05

Stephens and Donnelly(2000) introduced an efficient importance sampling scheme for computing the likelihood of a genetic sample. Their method consists of approximating the conditional probability of the allelic type of an additional gene given those currently in the sample. This technique was further extended by De Iorio and Griffiths (2004a,b) to a subdivided population framework for various mutation models between gene types. For sequence loci, the results of De Iorio and Griffiths (DIG) represent an improvement in terms of efficiency over those of Bahlo and Griffiths (2000) but haven't been tested beyond a single panmictic population. Using DIG's approach, we infer parameters from microsatellites evolving under a step-wise mutation model, DNA under an infinitely-many-sites model and single nucleotide polymorphisms (SNP) in a population of varying size. The inference from these markers is handled in a combined fashion by studying a single likelihood surface.

References:

Stephens, M., Donnelly, P., 2000. Inference in molecular population genetics. J. Roy. Statist. Soc. B 62, 605–655.
Bahlo, M., Griffiths, R.C., 2000. Inference from gene trees in a subdivided population. Theor. Popn. Biol. 57, 79-95.
De Iorio, M., Griffiths, R.C., 2004a. Importance sampling on coalescent histories, I. Adv. Appl. Probab. 36, 417–433.
De Iorio, M., Griffiths, R.C., 2004b. Importance sampling on coalescent histories, II. Subdivided population models. Adv. Appl. Probab. 36, 434-454.

A CONTEXT DEPENDENT PAIR HIDDEN MARKOV MODEL FOR STATISTICAL ALIGNMENT

Arribas-Gil, Ana ; Matias, Catherine
Departamento de Estadística, Universidad Carlos III de Madrid, Spain ; Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne, France

This work proposes a novel approach to statistical alignment of nucleotide sequences by introducing a context dependent structure on the substitution process in the underlying evolutionary model. We propose to estimate alignments and context dependent mutation rates relying on the observation of two homologous sequences. The procedure is based on a generalized pair-hidden Markov structure, where conditional on the alignment path, the nucleotide sequences follow a Markov distribution. We use a stochastic approximation expectation maximization (saem) algorithm to give accurate estimators of parameters and alignments. We provide results both on simulated data and vertebrate genomes, which are known to have a high mutation rate from CG

dinucleotide. In particular, we establish that the method improves the accuracy of the alignment of a human pseudogene and its functional gene.

ISOLATION-BY-DISTANCE MODEL REVISITED: ESTIMATING LOCAL AUTOCORRELATION PATTERNS

Duforêt-Frebourg Nicolas ; Michael G B Blum
Université Joseph Fourier, Centre National de la Recherche Scientifique, Laboratoire TIMC-IMAG UMR 5525, Grenoble, 38041, France
Under isolation-by-distance model, correlation between allele frequencies decays as a function of spatial distance. This is a pattern that is common for many species including humans. Here, we expand upon the traditional isolation-by-distance model and we propose a new statistical model where the spatial decay of correlation between allele frequencies may vary over space. Using Bayesian Gausian processes, we provide a spatial estimation of the decay parameter. Because the decay parameter will typically be increased in spatial regions with low gene flow, our new approach provides an alternative visualization of genetic structure for natural populations. Because the method is based on the correlation matrix between individuals or populations, it can scale to large data sets such as SNP chips. Using simulations that account for spatially varying migration/dispersion parameters, we show how our method compares to statistical multivariate approaches (PCA, MDS) and to the software BARRIER.

BUILDING A BIOINFORMATIC PIPELINE TO EXTIMATE BACTERIAL GENETIC DIVERSITY

Ferran Palero y Fernando González-Candelas
Unidad Mixta de Investigación en Genómica y Salud CSISP-Universitat de València
Molecular techniques have only recently started to unveil genetic diversity levels in bacterial species. Since its proposal by Maiden et al. (1998), Multilocus Sequence Typing (MLST) is the most generalised way of assessing molecular diversity of bacterial populations. Under the MLST scheme, isolates are characterized using the sequences of internal fragments (Approx. 450-500 bp) of six or seven house-keeping genes. The nucleotide differences between alleles are usually ignored, and each isolate of a species is characterised by a series of seven integers which correspond to the alleles at the seven house-keeping loci. A caveat that pervades most studies on bacterial population genetics is to which extent the MLST approach is providing an accurate representation of bacterial genetic diversity. The present work aims at defining how representative are the genetic diversity levels provided by the MLST house-keeping loci as compared with the global diversity levels provided by the whole of the genome.
In order to reach several public databases containing complete sequences of bacterial genomes and MLST isolate information, a computer pipeline was developed using PERL and Mathematica. Only bacterial species for which MLST schemes have been previously developed and for which more than five genome sequences are available were included in this study. For each species, all genomes were compared to build a pool of common genes or pangenome. Several summary statistics of nucleotide polymorphism levels were then estimated per gene, such as: the total number of segregating sites (S), the population mutational parameter (theta), and the nucleotide diversity (pi). Moreover, statistics for the neutrality-tests of Tajima (D) and Fu and Li (D* and F*) were calculated for each gene along the genome.
Our results show that genetic diversity estimates vary considerably along the genome of bacterial species. Interestingly, different levels of skewness for the distribution of diversity estimates and neutrality-test summary statistics were observed. In many cases, house-keeping loci showed lower genetic diversity levels than the average and presented neutrality-tests statistics far from zero, but the relative distance of the MLST loci to average value over the genome varied depending on the species.

A SITE-DEPENDENT EVOLUTIONARY MODEL FOR CRISPR SPACER ARRAYS

Kupczok Anne, Jonathan P. Bollback
IST Austria, Am Campus 1, 3400 Klosterneuburg, Austria
CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) consists of an array of repeats and spacers that acts as an adaptive heritable immune system found in Eubacteria and Archaea. The spacers between the repeats represent viral/plasmid targeting sequences and the system functions in an analogous way to the eukaryotic siRNA system. The length and content of the spacer array varies considerably among individuals within species (suggesting a rapid arms race) and it has been suggested that there is a selective cost, in the absence of parasites, associated with maintaining these arrays.
Therefore, the rate at which spacers are gained and lost from these arrays provides insight into the evolutionary dynamics of host-parasite interactions. To this end we develop a probabilistic model for the change of CRISPR spacer arrays over time. To describe biological observations, the model differs in two ways from standard phylogenetic models. First, insertion occurs only at the beginning of the array and second, one deletion event can affect multiple consecutive spacers. Parameter estimation under the model is done by maximum likelihood accounting for unobserved insertions and deletions. Simulating under this model shows that it differs considerably from site-independent models. The site-dependent and site-independent model are compared for bacterial CRISPR data sets.

- POSTERS (2) -

THE EVOLUTION OF GC CONTENT IN AVIAN GENOMES

Carina F. Mugal, Peter F. Arndt and Hans Ellegren
Department of Evolutionary Biology, Uppsala University

The genomes of many vertebrates, including mammals and birds, show a characteristic heterogeneous distribution of the local GC content, the so- called isochore structure of the genome. By now, the origin of isochores has been explained via the mechanism of GC-biased gene conversion (gBGC), i.e. short-scale, unidirectional exchanges between homologous chromosomes in the neighborhood of recombination-initiating double-strand breaks, where AT/GC heterozygotes produce more GC- than AT-gametes. Whereas the isochore structure is declining in many mammalian genomes, the heterogeneity in GC content is being reinforced in the avian genome. Despite this discrepancy, examinations of individual mammalian and avian substitution frequencies, are both consistent with the gBGC model of isochore evolution. However, a negative correlation between the local substitution rate and the local recombination rate present in the avian genome appears to be inconsistent with the gBGC model of evolution. Hence, it seems important to consider along with gBGC other consequences of recombination on the origin of mutations and their probability of fixation, as well as to take relationships of recombination rate to other genomic features into account.
In order to investigate the negative correlation between the local substitution rate and the local recombination rate, we developed a minimal analytical model to describe the substitution pattern found in the avian genome and compared simulated data to observed data. This analysis sheds light into which other genomic features and aspects of recombination impact on the local substitution pattern and the evolution of GC content in the avian genome. The results indicate that the local GC content itself, either directly or indirectly via interrelations to other genomic features, has an impact on the local substitution pattern by affecting the rate of mutation. Further, we suggest that this phenomenon is specific to avian genomes due to their unusually slow rate of chromosomal evolution, where many chromosomes have remained more or less intact during avian evolution. Because of this, interrelations between the local GC content and other genomic features are more pronounced in the avian genome.

THE EFFECT OF LONG BRANCHES ON MAXIMUM LIKELIHOOD TREE RECONSTRUCTION

Parks Sarah, Nick Goldman, EBI

Long branches, and the issues they cause during phylogeny reconstruction, have long been of interest in phylogenetics. Of particular concern has been long branch attraction (LBA), a regularly cited term generally used to describe a propensity for long branches to be placed near together in estimated trees. LBA was first mentioned in connection with inconsistency of parsimony, but it has since been claimed that the phenomenon exists for all major phylogenetic reconstruction methods. Despite the widespread use of this term in the literature, exactly what LBA is and what is causing it is poorly understood, even for simple evolutionary models and small model trees.
The statistical basis of ML, its robustness, and the fact that it appears to suffer less from biases lead to it being one of the most popular methods for tree reconstruction. However it is claimed that LBA does still occur, albeit in a less extreme way than for other methods. The widespread use of ML means that it is of interest to understand why, despite its statistical foundations, it may still suffer from LBA.
Studies looking at LBA have focused on the effect of two long branches on tree reconstruction. However, to understand the effect of two long branches it is also important to understand the effect of just one long branch. If ML struggles to reconstruct one long branch then this may have an impact on LBA. In this study we look at the effect of one long branch on three-taxon tree reconstruction. We show that the results are counterintuitive but can be understood through the use of analytical solutions to the ML equation and distance matrix methods. We build upon these results to look at the effect of two long branches on four-taxon trees, considering both the different topologies inferred and the placement of the long branches on the different topologies.
The results to be presented illustrate that even small model trees are still interesting to help understand how ML phylogenetic reconstruction works and that LBA is a complicated phenomenon that deserves further study.

REVBAYES: ONE PROGRAM FOR YOUR WHOLE PHYLOGENETIC ANALYSIS

Höhna Sebastian, Fredrik Ronquist, John Huelsenbeck
Museum of Natural History, Stockholm, Sweden; University of California, Berkeley, Ca, USA

The amount of models has increased rapidly in recent years. The models used in molecular evolutionary analyses range from the gene level (e.g. substitution models) to the species level (e.g. diversification models and phylogeographic models). This plethora of models has retained a rich landscape of computer programs where most models are only implemented in one single program. Hence, most analyses are performed gradually using different computer programs. However, there are several reasons why a single software unifying as many models as possible is beneficial:
First, an integrative analysis can combine several models and therefore take the uncertainty in the different steps of the analysis into account. Second, model selection and model choice is in most situations only feasible if competing models are implemented in the same computer program. Furthermore, mixtures models are easier to implement if more models are available.

Our new computer program, RevBayes (expected release: June 2012), is an R-like environment for Bayesian phylogenetic inference which aims to integrate the whole analysis. Most standard models are implemented and it already offers more models than its predecessor MrBayes. Even if a model is missing, adding the new model is as easy as writing a function in R. RevBayes provides a very flexible environment to specify a model and run a phylogenetic analysis. Once a model is specified, RevBayes offers not only the inference machinery but also simulation tools for observations under the model.
In this talk I will give a brief overview of RevBayes and demonstrate its abilities on an integrative, fully hierarchical Bayesian analysis where, amongst others, the tree topology, the divergence times, clock rates and diversification rates are estimated simultaneously. Several models and mixture thereof are considered and evaluated using marginal likelihoods. Our preliminary results show that the tree topology, divergence times and diversifications rates can be biased if estimated independently.

A DIVIDE AND CONCATENATION STRATEGY FOR THE PHYLOGENETIC RECONSTRUCTION OF LARGE ORTHOLOGOUS DATASETS

Chang Jia-Ming(1), Matthieu Muffato (2); Jean-francois Taly (1); Javier Herrero (2); Cedric Notredame (1)
(1) Centre for Genomic Regulation (CRG) and UPF, Barcelona, Spain; (2) European Bioinformatics Institute, United Kingdom
Thanks to next-generation sequencing technique, more and more sequences have become available. This overwhelming amount of data is challenging even the fastest methods and some important multi-genetic families like olfactory receptors have become impossible to analyze with the most accurate methods like Maximum Likelihood.
Here we show how a simple Divide and Conquer (D&C) strategy can be applied to this problem by breaking it down in smaller problems, that are solved independently. Our approach relies on the ability to identify within a large dataset clusters of orthologous genes. Each cluster of orthologous genes is used to build a phylogenetic tree using a typical approach (M-Coffee + TreeBest). The upper level of the tree (super-tree) is resolved in a second stage. One protein per species is chosen from each subtree. All proteins from the same species are aligned together. The alignment used for building the super-tree results from concatenating all these alignments. The advantage is that we reduce the number of sequences to classify without loosing information as all sequences are represented in the final alignment. This approach can easily deal with lineage specific duplications, but not with lateral transfers. It is therefore better suited for the analysis of eukaryotic large families like the kinases or the olfactory receptors.
We applied this approach for classifying 858 olfactory receptors from 13 Drosophila species. We could reconstruct the tree of all these sequences (403 amino acids long on average) in roughly one hour in a normal workstation. The log likelihood of the final tree is -739,510, superior to -1,381,448 with the traditional neighbor joining (NJ) method and is quite close to optimal -722,299 (by PhyML, 96h7m22s).

ARMADILLO 1.1: AN ORIGINAL WORKFLOW PLATFORM FOR DESIGNING AND CONDUCTING PHYLOGENETIC ANALYSIS AND SIMULATIONS

Diallo Abdoulaye Baniré, Etienne Lord, Mickael Leclercq, Alix Boc and Vladimir Makarenkov
201, avenue du Président-Kennedy
I will present how our new framework Armadillo v1.1 can easily fit large scale analysis of different phylogenetics recontruction. This platform is a novel workflow platform dedicated to designing and conducting phylogenetic studies, including comprehensive simulations. A number of important phylogenetic and general bioinformatics tools have been included in the first software release. As Armadillo is an open-source project, it allows scientists to develop their own modules as well as to integrate existing computer applications. Using our workflow platform, different complex phylogenetic tasks can be modeled and presented in a single workflow without any prior knowledge of programming techniques.
I will also present how we added a machine learning techniques to guide on the good choice of methologic to solve a specific problem via a tutoring intelligent system. The program and its source code are freely available at: <http://www.bioinfo.uqam.ca/armadillo>.

IS THE PROTEIN MODEL ASSIGNMENT PROBLEM NP-HARD?

Kassian Kobert, Alexandros Stamatakis; Jörg Hauser
HITS gGmbH, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
In phylogenetics, computing the likelihood that a given tree generated the sequence data at hand requires calculating the probability of the available data under a given phylogeny and a statistical model of sequence evolution.
Here, we focus on how to select an appropriate model for the data at hand which represents an important and generally non-trivial task. It is well-known that, an inappropriate model which does not fit the data well, can generate erroneous phylogenetic estimates.
More specifically, we consider the case of partitioned protein sequence alignments for which we intend to assign an individual, best-fit, protein substitution matrix to each partition (e.g., each gene), while branch lengths are linked across partitions (joint branch length estimate) to reduce the number of free parameters in the model.
Note that, each set of independent per-partition branch lengths increases the number of model parameters by 2n-3 where n is the number of taxa.
For instance, when analyzing large multi-gene datasets it may be appropriate to link certain partitions/genes via a joint branch length estimate to reduce the number of model parameters and avoid over-parameterizing the model.
Our objective is thus to maximize the log likelihood score for the per-partition protein model assignments (e.g., JTT, WAG, LG, etc) under a joint branch length estimate on a given, fixed, and reasonable (i.e., non-random) tree topology.

Simply comparing the log likelihood scores of all possible model assignments quickly becomes computationally prohibitive because of the exponential number of possible model combinations that need to be computed.
In fact, we show that, the problem of finding the optimal protein substitution model configuration under linked branch lengths on a given, fixed tree, is NP-hard.
That is, unless P=NP no efficient algorithm exists that guarantees an optimal solution in polynomial time for the protein model assignment problem as defined here.
Thus, our result implies that, one should rather employ heuristics that can approximate the solution, than striving to find the exact solution for a large number of partitions.
Alternatively, the problem can be simplified by relaxing the assumptions. For instance, one can unlink the branch lengths or reduce the number of parameters by conducting a joint branch length estimate for a sufficiently small subset of partitions such that the optimal solution can be computed by means of an exhaustive search.

EVOLUTIONARY COSTS IN GENE-SPECIES RECONCILIATION

Pawel Gorecki (2), Oliver Eulenstein (1), Jerzy Tiuryn (2)
(1) Iowa State University, Ames IA, USA, (2) University of Warsaw, Institute of Informatics, Warsaw, Poland
During this talk I will present our recent results on the process of reconciling of an unrooted gene tree and rooted species tree under several known evolutionary costs and distances, such as, duplication-losses, deep coalescence, losses, duplications and Robinson-Foulds. We focus on mathematical and algorithmic properties of the problems related to this model: rooting of a gene tree, error correction of gene trees, local search and supertree problems.

ANISOTROPIC ISOLATION BY DISTANCE: THE MAIN ORIENTATIONS OF HUMAN GENETIC DIFFERENTIATION

Flora Jay, Per Sjödin, Mattias Jakobsson , Michael G. B. Blum
Université Joseph Fourier, Grenoble; CNRS; University of California, Berkeley, Uppsala University
Genetic differentiation between human populations is greatly influenced by geography because of the accumulation of local genetic differences. However, there has been no attempt so far at describing the different increment of genetic differentiation along the different azimuthal orientations. We analyzed genome-wide polymorphism data from African (n=29), Asiatic (n=26), Native American (n=9) and European (n=38) populations and we found that the major orientations of genetic differentiation are north-south in Europe and Africa, east-west in Asia whereas we did not find an anisotropic pattern in the Americas. A practical consequence of the anisotropic pattern of genetic differentiation is that the localization of an individual's geographic origin based on his SNP data should be facilitated in the orientation of maximum differentiation. We compared localization of geographic origin obtained with principal component regression and with a baseline method and we confirmed that the largest improvement was obtained in the orientation of maximum differentiation. Our findings have implications for interpreting the making of current human gene pool in terms of isolation by distance and spatial range expansion processes.

INTERACTIONS BETWEEN MUTATIONS AND MAINTENANCE OF SEX IN SUBDIVIDED POPULATIONS.

Matthew Hartfield (1), Sarah P. Otto (2), Peter D. Keightley (1).
(1) Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom
(2) Department of Zoology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada
Although there is no known general explanation as to why sexual populations resist asexual invasion, previous work has shown that sexuals can outcompete asexuals in structured populations. However, it is currently unknown whether costly sex can be maintained with weak structure that is commonly observed in nature. We investigate the conditions under which obligate sexuals resist asexual invasion in structured populations subject to recurrent mutation. We determine the level of population structure needed to disfavour asexuals, as calculated using the average Fst between all pairs of demes. We show that the critical Fst needed to maintain sex decreases as the population size increases, and approaches modest levels as observed in many natural populations. Sex is maintained with lower Fst if mutation is both advantageous and deleterious, if mutation rates are sufficiently high, and if deleterious mutants have intermediate selective strengths, which maximises the effect of Muller's Ratchet. Additionally, the critical Fst needed to maintain sex is lower when there are a large number of subpopulations. Lower Fst values are needed to maintain sex when demes vary substantially in their pairwise distances (e.g., when arrayed along one dimension), although this effect is often modest, especially if some long-distance dispersal is present.

CORRECTING PRINCIPAL COMPONENTS OF SPATIAL POPULATION GENETIC VARIATION UNDER ISOLATION BY DISTANCE MODELS

E. Frichot(1), S. Schoville(1), G. Bouchard(2), O. François(1)
(1) TIMC-IMAG, Grenoble, France (2) XRCE, Grenoble, France
In many species, spatial population genetic variation displays patterns of isolation by distance. Characterized by locally correlated allele frequencies, these patterns are known to create periodic shapes in geographic maps of principal components which confound signatures of specific migration events and influence interpretations of principal component analyses (PCA). In this presentation, we introduce a new model combining probabilistic PCA and Bayesian kriging to infer population genetic structure from spatial genetic data while correcting for errors introduced by isolation by distance. The proposed algorithm is based on matrix factorization and low rank approximations, and scales with the dimension of the data set. To illustrate our method, we generated patterns of isolation by distance and broad-scale geographic clines using simulations of spatial Markov models. We show that our method improves the

interpretation of PC maps, and is able to remove the horseshoe patterns usually observed in those maps for spatially correlated data. We present an application to human SNP data from the Human Genome Diversity Panel.

AN EVOLUTION MODEL FOR SEQUENCE LENGTH BASED ON RESIDUE INSERTION-DELETION INDEPENDENT OF SUBSTITUTION: AN APPLICATION TO THE GC CONTENT IN BACTERIAL GENOMES

S. Lèbre, C. Michel

We introduce here a gene evolution model which is an extension of the time-continuous stochastic IDIS model [Lèbre and Michel, 2010] to sequence length. In the IDIS model, the insertion rates are explicit parameters which are entirely independent from the substitution parameters. Let us consider an alphabet of K residues. Let ni(t) be the occurrence number of residue i in the biological sequence at time t. Let ri be the insertion rate per site of each residue i, 81 <= i <= K, ri >= 0. Following a classical model in population dynamics [Malthus, 2000], we assume that the growth rate n'i(t) = dni(t)/dt of each residue i at time t due to insertions is equal to ri x n(t). Let d be the deletion rate for all residues, d >=0. The growth rate n'i(t) of ecah residue i at time t due to deletions is d x ni(t) where ni(t) is the number of occurrences of residue i in the sequence, then for all 1 <= i <= K, n'i(t) = ri x n(t) - d x ni(t).

We derive a general matrix differential equation allowing for the substitution process and the insertion-deletion process to be superimposed where these processes are assumed to be independent, i.e. a substitution event does not alter the probability of an insertion-deletion event and reciprocally. Then the derivative P'(t) of the residue occurrence probability at time t is the result of the instantaneous variation due to the substitution. The insertion-deletion and the occurrence probability vector P(t) = [Pi(t)]1<=i<=K satisfies the following nonhomogeneous matrix linear differential equation P'(t) = A . P(t) + R where A = M - (1 + r) I, M is the substitution probability matrix (stochastic in column), R = [ri]1<=i<=K is the vector of the residue insertion rates per site and T = sum 1<=i<=K ri is the sum of the residue insertion rates.

The novel IDISL (Insertion Deletion Independent of Substitution based on sequence Length) model gives an analytical expression of the residue occurrence probability p(l) at sequence length l = n(t) observed at time t, depending on stochastically independent processes of substitution, insertion and deletion [Lèbre and Michel, 2012]. For any diagonalizable substitution matrix M, the residue occurrence probability p(l) is given as a function of the eigenvalues (Lk)1<=k<=K of M, the eigenvector matrix Q of M, a vector r of the residue insertion rates, a deletion rate d (unlike our previous IDIS model) and a vector of the initial residue occurrence probability p(lo) at sequence length lo (…).

The IDIS class of models allows a mathematical analysis of the behavior of the residue occurrence probability according to either evolution time or sequence length. The length parameter can be associated with any nucleotide regions: genes, genomes, introns, repeats, 50 and 30 regions, etc. Three properties of the IDISL model are given in relation with the sequence length l: parameter scale, inverse evolution and residue equilibrium distribution. As an illustration, nucleotide occurrence probabilities are given in the particular case of the IDISL-HKY model, i.e. the IDISL model associated with the HKY asymmetric substitution matrix [Hasegawa et al., 1985]. An application of the IDISL model with HKY substitution matrix is developed for a massive statistical analysis of GC content in all complete bacterial genomes available to date (NCBI, 894 non-anaerobic and anaerobic genomes). The IDISL model confirms the increase of the GC content with the genome length for two non-anaerobic taxonomic groups of bacterial genomes. Moreover, the non-linear modelling proposed by the IDISL model outperforms the most recent modelling of GC content in these bacterial genomes.

SUPERQ: A NEW METHOD TO CONSTRUCT WEIGHTED SUPERNETWORKS FROM PARTIAL TREES

Sarah Bastkowski, University of East Anglia

Presenting evolutionary data from different trees for a set of taxa in a joint network is an important problem in phylogenetics. A tree construction for one single gene does not always correspond to the species phylogeny. So a common approach is to sequence many genes, construct trees for each of them and fuse the trees into a supertree or a network. This becomes even more difficult if the input consists of partial trees, i.e. several taxa are missing in these gene trees. In the SuperQ algorithm, quartets, which are derived from the input trees, are used to find a circular split network. The next challenge is to find good weights for the splits in the resulting network. We describe a new approach to this problem and present some results on the performance of SuperQ.