Mathematics and Informatics in Evolution and Phylogeny

June 10-12, 2008

Hameau de l'Etoile

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ MONTPELLIER 2 SCIENCES ET TECHNIQUES

Méthodes et Algorithmes pour la Bio-informatique LIRMM

GdR BiM

la Région Languedoc Roussillon

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTERE DES AFFAIRES ETRANGERES ET EUROPEENNES

Tuesday, June 10th

9h **Bus from Montpellier centre to Hameau de l'Etoile**

10h **Café & croissants**

10h50 **Bienvenue**

11h – 12h15 **Information Theory**

11h ***Beyond 'event horizons' in early evolution***
Mike Steel, University of Canterbury, New Zealand.

11h25 ***An Equivalence of Maximum Parsimony and Maximum Likelihood revisited***
Mareike Fischer, University of Canterbury, New Zealand.

11h50 ***Identifiability of models from parsimony-informative pattern frequencies***
John Rhodes, University of Alsaska Fairbanks, USA.

12h15 **Déjeuner**

13h45 – 15h25 **Models**

13h45 ***Realism and Instrumentalism in theoretical models of molecular evolution***
David Penny, Massey University, New Zealand.

14h10 ***Probabilistic models of the evolution of the rate of evolution***
Stéphane Guindon, University of Auckland, New Zealand.

14h35 ***Quantifying the Time Irreversibility of the Nucleotide Substitution Process***
Federico Squartini, Max Planck Institute, Germany.

15h ***Estimating the contribution of sequence context to nucleotide substitution rate heterogeneity***
Helen Lindsay, Australian National University, Australia.

15h25 – 16h **Thé & gateaux**

16h – 17h45 **Models & Algorithms**

16h ***Using the best model! Reconstructing the worst tree…***
Liat Shavit-Grievink, Massey University, New Zealand.

16h10 ***Awesome Matrices and Markov models in Phylogenetics***
Steffen Klaere, Center for Integrative Bioinformatics Vienna , Austria.

16h20 ***A new phylo-HMM paradigm to search for sequences***
Jean-Baka Domelevo-Entfellner, LIRMM – CNRS, France.

16h30 ***Modelling heterogeneity in nucleotide sequence evolution***
Simon Whelan, University of Manchester, UK.

16h55 ***Sampling trees with a fixed number of leaves***
Tanja Gernhard, Technische Universität München, Germany.

17h20 ***Simultaneous estimation of alignments and trees***
Tandy Warnow, University of Texas, USA.

Wednesday, June 11th

9h – 10h15 **Rearrangements & duplications**

9h ***Including descendants of whole genome duplication in gene order phylogeny***
David Sankoff, University of Ottawa, Canada.

9h25 ***Reconstruction of ancestral chromosomes: methodological frameworks***
Eric Tannier, LBBE – INRIA, France.

9h50 ***The complexity of deriving multi-labeled trees from bipartitions***
Vincent Moulton, University of East Anglia, UK.

10h15 – 10h45 **Café & croissants**

10h45 – 12h25 **Bayesian**

10h45 ***Comparison of uncorrelated and autocorrelated Relaxed Phylogenetics***
Michael Defoin-Platel, University of Auckland, New Zealand.

11h15 ***Bayesian estimation of selection pressure on protein coding sequences***
Aude Grelaud, MIG – INRA, France.

11h35 ***The dynamics of positive selection on the mammalian tree: A Bayesian inference of selection histories***
Carolin Kosiol, Cornell University , USA.

12h ***Studying historic demographic parameters using an approximate bayesian computation***
Joao Lopes, University of Reading, UK.

12h25 **Déjeuner**

13h45 – 14h50 **Genetics**

13h45 ***Evolutionary pathways for sex-determining mechanisms based on X-chromosome elimination: The sciarid system***
Lucas Sanchez, CSIC, Spain.

13h55 ***Evolutionary process of a tetranucleotide microsatellite locus in Acipenseriforms***
Eric Rivals, LIRMM – CNRS, France.

14h05 ***Application of Matlab in population genetics and molecular evolution***
James Cai, Stanford University, USA.

14h15 ***Expected time to coalescence and Fst under a skewed offspring distribution among individuals in a population***
Bjarki Eldon, Harvard University, USA.

14h25 ***Coalescence: exact calculations of coalescence times for subdivided population***
Alain Franc, BGC – INRA, France.

14h50 – 15h10 **Pause & thé**

15h10 – 16h **Networks**

15h10 ***Summarizing Multiple Gene Trees Using Cluster Networks***
Regula Rupp, University of Tübingen, Germany.

15h35 ***Level-k phylogenetic networks***
Leo van Iersel, Techinische Universiteit Eindhoven, Netherlands.

16h **Thé & gateaux, promenades, discussions …**

Thursday, June 12th

9h – 10h40 **Trees & supertrees**

9h ***Phylogenetic Diversity with Disappearing Features***
Charles Semple, University of Canterbury, New Zealand.

9h25 ***Encoding phylogenetic trees in terms of weighted quartets***
Katharine Huber, University of East Anglia, UK.

9h50 ***Healing source trees to obtain healthy supertrees***
Céline Scornavacca, LIRMM – CNRS, France.

10h15 ***Human genetic ancestry: When branches get short***
Arndt Haeseler, Center for Integrative Bioinformatics Vienna, Austria.

10h40 – 11h10 **Café & croissants**

11h10 – 12h25 **Viruses**

11h10 ***The computational biology of genetically diverse assemblages***
Allen Rodrigo, University of Auckland, New Zealand.

11h35 ***Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus***
Gaël Thébaud, BGPI – INRA, France.

12h ***Applying molecular clocks to date the intraspecific and interspecific diversification of RNA plant virus species***
Denis Fargette, RPB – IRD, France.

12h25 **Déjeuner**

13h45 – 15h40 **Phylogenetics**

13h45 ***A phylogenetic follow-up study of 4 individuals infected with closely related HIV-1 strains***
Kristen Chalmet, Ghent University , Belgium.

13h55 ***Comparison of commonly used methods for combining multiple phylogenetic data sets***
Anne Kupczok, Center for Integrative Bioinformatics Vienna, Austria.

14h05 ***Phylogenetic Diversity with Ecological Constraints***
Beáta Faller, University of Canterbury, New Zealand.

14h15 ***Tools for choosing among alternative network phylogenetic inferences***
Steven M. Woolley, St. Louis, USA.

14h25 ***Parallel Adaptations to High Temperatures in the Archean Eon***
Samuel Blanquart, LIRMM – CNRS, France.

14h50 – 15h10 **Pause & thé**

15h10 – 16h25 **Models & Algorithms**

15h10 ***Is protein sequence evolution constant over time?***
Nick Goldman, EMBL-European Bioinformatics Institute, UK.

15h35 ***New Methods for Genealogical Network Inference based on Local Tree Topologies with a Set of SNP Sequences in Populations***
Yufeng Wu, University of Connecticut, USA.

16h ***Exact and efficient algorithms for the probability of a marker under incomplete lineage sorting***
David Bryant, University of Auckland, New Zealand.

16h25 **The end! Bus to Montpellier centre**

Mathematics and Informatics in Evolution and Phylogeny

June 10-12, 2008
Hameau de l'Etoile, Montpellier, France

Theme

The subject is evolution, which is considered at different scales: sequences, genes, gene families, organelles, genomes, and species. The focus is on the mathematical and computational tools and concepts, which form an essential basis of evolutionary studies. Recent years have witnessed rapid progress in this area, with models becoming more realistic, and complex, and with fast algorithms able to deal with the large datasets that are available today. The main topics are: phylogenetics, evolutionary genetics and genomics, molecular evolution of pathogens and epidemiology, biodiversity, statistical modelling, algorithmics, and software development. The programme includes short (25 minutes) and flash (10 minutes) talks, and plenty of time for discussions.

Scientific Committee

Elisabeth Allman: University of Alaska, US.
Vincent Berry: LIRMM, Université de Montpellier, FR.
David Bryant: University of Auckland, NZ.
Frantz Depaulis: ENS, CNRS, FR.
Laurent Duret: BBE, CNRS, FR.
Nicolas Galtier: ISEM, CNRS, FR.
Olivier Gascuel: LIRMM, CNRS, FR, chair.
Junhyong Kim: University of Pennsylvania, US.
Mike Hendy: Massey University, NZ.
Daniel Huson: University of Tübingen, DE.
Vincent Moulton: University of East Anglia, UK.
David Posada: Universidad de Vigo, ES.
Allen Rodrigo: University of Auckland, NZ, co-chair.
Noah Rosenberg: University of Michigan, US.
Charles Semple: University of Canterbury, NZ.
Mike Steel: University of Canterbury, NZ.

Jules-Sébastien-César Dumont D'Urville

Organizing Committee

Céline Berger, Samuel Blanquart, Olivier Gascuel, Vincent Lefort (LIRMM, CNRS-UM2)

Funding

This conference is a continuation of the first Dumont D'Urville Workshop on Applied Evolutionary Bioinformatics, which was held at the University of Canterbury in New Zealand in June 2007. As the first edition, this conference is supported by the MAE (France) and MORST (NZ), through the Dumont D'Urville programme, to promote and support scientific cooperation between New Zealand and French researchers. Moreover, the conference is largely open to attendees from other countries, and we received funding from the University of Montpellier II, the Languedoc-Roussillon region, and the GDR Bioinformatique Moléculaire (CNRS). We would like to thank all these sponsors for their help.

Beyond 'event horizons' in early evolution

Mike Steel, Elchanan Mossel, Laszlo Szekely

Biomathematics Research Centre, University of Canterbury, New Zealand.

How far back in time can one reliably reconstruct phylogenetic history? Much depends on substitution rates, relative branch lengths and properties of the underlying evolutionary model. Perhaps some early divergence events are fundamentally 'unresolvable' even if we had all the genomic data present in all extant organisms (and knew what to do with it). This talk will outline some ways in which probability theory can determine the phylogenetic information content of genetic sequences.

An Equivalence of Maximum Parsimony and Maximum Likelihood revisited

Mareike Fischer, Bhalchandra D. Thatte

Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand.

The incessantly growing amount of available genetic sequence data requires both stochastic models for nucleotide substition as well as tree reconstruction methods to allow for the inference of phylogenetic trees. Unsurprisingly, such models and methods have therefore been widely discussed in the last decades. Two of the most frequently used tree reconstruction methods are Maximum Parsimony (MP) and Maximum Likelihood (ML), and it is known that these methods sometimes disagree (e.g. in the so-called Felsenstein zone). However, in 1997 Tuffley and Steel proved that under a symmetric multistate model of the evolutionary process, when applied to a sequence of no common mechanism, MP and ML are equivalent.

In my talk, I will present an elementary new proof for this result. Moreover, I will show that small changes of the model assumptions suffice to cause MP and ML to disagree even when the sequence data evolved under no common mechanism. In particular, I will present an example in which the probabilities of substitutions are small and MP and ML are not equivalent. This result is particularly surprising as MP normally is assumed to be justified whenever mutation events are rare.

Identifiability of models from parsimony-informative pattern frequencies

John Rhodes[1], Elizabeth Allman[1], Mark Holder[2]

[1]University of Alaska Fairbanks, USA.
[2]University of Kansas, USA.

When morphological or other non-sequence data is collected for phylogenetic inference, there is an acquisition bias towards collection of parsimony-informative characters. Lewis (2001) proposed the use of model-based Maximum Likelihood inference for morphological data, by using likelihoods conditioned on characters being nonconstant. The models underlying this are known to be identifiable, essentially as a corollary to the identifiability of models with invariable sites. This identifiability ensures statistical consistency of ML inference.

Autapomorphic (nonconstant, but parsimony-noninformative) characters are also likely to be subject to substantial acquisition bias, however, and are often missing from data sets collected with parsimony analysis in mind. We investigate the identifiability of simple models from only parsimony-informative information, obtaining both positive and negative results.

Realism and Instrumentalism in theoretical models of molecular evolution

David Penny

Institute of Molecular, BioSciences, Massey University, Private Bag 11222, Palmerston North, New Zealand.

There have always been at least two themes or approaches to scientific models. For example, Copernicus's' model was considered okay by the authorities if is was only considered an 'instrumentalist' model - a useful instrument for calculating import events in the Calendar, but unacceptable of it was considered a 'real' model of the solar system. In models of molecular evolution we have the same two approaches occurring. One is to have models with the smallest number of parameters, to help both statistical power and identifiability. However, the parameters may have to make assumptions that are known to be incorrect biochemically. Conversely, biochemically realistic models may have many more parameters, and lead to apoplexy amongst purists. Both approaches have their scientific utility, and perhaps we should focus more on the biological question that is being addressed, including conditional probabilities and time periods over which an event may have occurred - not just point estimates of times of divergence.

Probabilistic models of the evolution of the rate of evolution

Stéphane Guindon

Department of Statistics, University of Auckland, New Zealand.
Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, France.

New models that describe the evolution of the substitution rates during the course of evolution will be introduced. The rate of substitution is considered here as a stochastic process which statistical properties will be detailed and discussed in the light of biological knowledge. The combination of these models with the estimation of node times in a phylogeny defines a new and suitable approach for molecular dating in a sound statistical framework.

Quantifying the Time Irreversibility of the Nucleotide Substitution Process

Federico Squartini and Peter F. Arndt

Max Planck Institute for Molecular Genetics, Department of Computational Molecular Biology, Ihnestr, 63/73, 14195 Berlin, Germany.

Markov models of genome evolution, widely used in phylogeny reconstruction, usually assume the time reversibility of the nucleotide substitution process. Although these models give meaningful results when applied to biological data, it is not clear if the assumption of time reversibility really holds and, if not, how much sequence evolution processes deviate from it. To this aim we have introduced the irreversibility index (IRI), deriving it from the Kolmogorov cycle conditions for time reversibility (an alternative to the detailed balance condition).

The IRI is a function of the rate matrix and is a direct measure of the degree of time-reversibility of the evolutionary nucleotide substitution process. We first derive the IRI for Markov models describing the evolution of independent nucleotide sites. Since neighbor dependencies play a significant role in the evolution of vertebrate genomes (due to the CpG methylation-deamination process), we extended the index to models of evolution which also take into account nearest-neighbor dependencies along the nucleotide sequence.

Obviously, in order to compute the IRI for the evolution of a given species lineage, it is necessary to infer the evolutionary rates in a framework which does not assume time reversibility a priori. We show how this can be done using a mixed Monte-Carlo Maximum-Likelihood (MCML) approach, which combines elements of the two methods in a very efficient way. The method requires

a three species alignment (two sister species and an outgroup) and relies on the reconstruction of the ancestral sequence of the one internal node of the phylogeny.

As an application of the above formalism, we have computed the IRI for the evolutionary processes of two lineages, Drosophila simulans and Homo sapiens. In the first case we have disregarded neighbour dependencies and used the simpler form of the IRI. In the case of the human genome instead, we have also taken into account the CpG decay process. As a result of the analysis we have found in both cases statistically significant deviations from the ideal case of time reversibility.

Estimating the contribution of sequence context to nucleotide substitution rate heterogeneity

Helen Lindsay, Gavin A. Huttley

John Curtin School of Medical Research, Australian National University, Acton, ACT, 0200, Australia.

Nucleotide substitution rate heterogeneity is well known and has traditionally been modelled using a gamma distribution that assigns nucleotide sites to different rate classes. An alternative is that substitution rate heterogeneity may reflect the effects of local sequence context on nucleotide substitution rates. Under a gamma model, variations in the rate of a given type of nucleotide substitution arise from site-specific constraints. At a fast evolving site, all types of substitutions occur rapidly compared to the average rate. Under a context-dependent substitution process, the substitution rate of a site depends on its current neighbours and may change though time according to whether the current local sequence context promotes or inhibits nucleotide substitution. If substitution rate heterogeneity reflects context dependent processes, modelling this effect using a gamma rate parameter could potentially result in statistical support for an incorrect tree. The evolutionary importance of sequence context effects, and the extent to which substitution rate variation between sites reflects context effects are unknown. We have implemented novel dinucleotide substitution models to disentangle context-dependent and independent substitution patterns using the PyCogent software package. Considering equivalently parameter-rich models applied to a primate data set consisting of human, chimpanzee and macaque introns, we find that our dinucleotide models are approximately five times better able to describe the substitution process than nucleotide substitution models incorporating gamma rate heterogeneity. The improvements derived by including context-dependent and gamma parameters are non-additive, indicating that the gamma parameter partially captures effects arising from local sequence context. Our analyses suggest that models which allow for substitution rates to vary through time as well as between sites

are needed for accurate phylogenetic reconstruction.

Using the best model! Reconstructing the worst tree

L. Shavit-Grievink, B. Holland, M. Hendy, P. Lockhart and D. Penny

Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, New Zealand.

Commonly used phylogenetic models assume a homogeneous process through time in all parts of the tree; they do not account for lineage specific properties. However it is known that these models are too simplistic, and with time the processes of evolution can change. In particular, it is now widely recognized that as constraints on sequences evolve the proportion of variable sites can vary between lineages. This will affect the ability of phylogenetic methods to correctly estimate phylogenetic trees, especially for long timescales. To date there is no phylogenetic model that allows for change in proportion of variable sites, and the degree to which this affects phylogenetic reconstruction is still unknown. We present a modification to the program seqgen-cov that allows the generation of sequences with changes in proportion of variable sites. We used this simulator to evaluate Bayesian tree reconstruction when the proportion of variable sites changes through time. We show that tree reconstruction under the best fitting homogeneous, time reversible, stationary model often results in the wrong tree.

Awesome Matrices and Markov models in Phylogenetics

Tanja Gesell, Arndt von Haeseler, and **Steffen Klaere**

Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University Vienna, Veterinary University Vienna, Austria.

Usually, models of evolution evaluate the changes of sequences or single nucleotides through mutations along a tree. We apply these models to the space of site patterns. In particular, we present tree-related one step mutation matrices (OSM) which describe the change of patterns at the leaves of a tree when one mutation occurs on the tree.

These matrices are constructed as convex sums of permutation matrices (with each identifying mutations on a branch). They have a lot of nice computational properties making various calculations quite simple. Moreover, these

matrices incorporate maximum parsimony, maximum likelihood and provide other tree-related measures under one framework thus giving us the opportunity to compare these methods and to further analyze certain tree-related events like the Felsenstein zone.

We will present the framework on the general two-state model and introduce the various tree-related methods. We will also present an extension to the Kimura 3st model and the problems extending the framework to more general models.

A new phylo-HMM paradigm to search for sequences

Jean-Baka Domelevo Entfellner, Olivier Gascuel

Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR 5506, CNRS-Université de Montpellier 2, 161 rue Ada, 34392 Montpellier Cedex 5, France.

We introduce a new type of phylogenetic Hidden Markov Model, combining the strength of usual HMM and the knowledge of the phylogeny of a family of sequences. We use such models to look into the genome of a target species for homologs to the aligned proteins. Our results on some 700 protein families show a better sensitivity and a better specificity when compared to standard profile HMM.

Modelling heterogeneity in nucleotide sequence evolution

Simon Whelan

University of Manchester, UK.

Modern phylogenetics depends on models that describe how nucleotides and amino acids substitute each other during evolution. These models typically make many simplifying assumptions about the evolutionary process, including that evolution looks the same at each site in a sequence and on all the branches of a phylogenetic tree. Many studies have shown that in reality evolution is heterogeneous and that failure to account for this variability can introduce systematic errors into many forms of phylogenetic analysis. Current models struggle to account for this heterogeneity, with the complexity of the model often growing to an unacceptable degree as the amount of data grows. I introduce a simple modelling approach that can incorporate complex patterns of heterogeneity in rate, nucleotide frequencies, and substitution bias. I proceed to demonstrate

how this model can be used to characterise different types of heterogeneity in coding sequences, and investigate its ability to correct for complex dependencies between sites in sequence data.

Sampling trees with a fixed number of leaves

Tanja Gernhard[1], Klaas Hartmann[2], Dennis Wong[3]

[1]Zentrum Mathematik (M9), Technische Universität München, Germany.
[2]University of Canterbury, New Zealand.
[3]New Brunswick University, Canada.

Various stochastic models have been proposed for the process of speciation. In order to understand the models, sampling trees from the model distribution is a crucial task. Sampling trees with a fixed age is straightforward – run a simulation until the age is reached. However, in order to compare a reconstructed phylogeny with the model, we need to sample trees with a fixed number n of leaves. This can be a tricky task even for pure birth models – we will explain why simulating until n species are obtained is wrong for general (pure birth) models. We will provide algorithms which sample trees for any speciation model correctly. For certain classes of models, we use analytic results for obtaining simpler and faster sampling methods. The methods presented are all implemented in Perl.

Simultaneous estimation of alignments and trees

T. Warnow, C.R. Linder, K. Liu, S. Nelesen

The University of Texas at Austin, Austin TX 78746, USA.

Biomolecular sequences evolve under processes that include substitutions as well as more complex events, such as insertions and deletions of subsequences, duplications of subsequences, and rearrangements. As a result, the full phylogenetic reconstruction of a collection of sequences not only produces a hypothesis of the evolutionary tree (or network) underlying the data, but also a hypothesis of the events that were involved in producing the observed data.

Consequently, a phylogenetic estimation will generally include the estimation of the multiple alignment on the sequences, as well as the phylogeny. Most phylogenetic analyses separate these operations into two phases: first estimating the multiple sequence alignment (MSA) and then estimating the tree based upon that alignment. However, simultaneous estimation of trees and alignments has also been attempted, with some newer methods based upon likelihood, and some (e.g., POY) based upon extensions of parsimony to include gaps in the calculation of treelength.

In this talk, we will present the results of our simulation studies that show that the existing methods all fail to produce good estimates of the true evolutionary tree under conditions that include many taxa evolving under even moderately high rates of indels and site substitutions. We will also describe our new method "SATe" (Simultaneous Alignment and Tree Estimation) and present results showing how it improves upon existing methods. Finally, we will present initial results of our most recent work in which we attempt to solve maximum likelihood under models that include gaps.

Including descendants of whole genome duplication in gene order phylogeny

Chunfang Zheng, Qian Zhu, **David Sankoff**

University of Ottawa, Canada.

Basic rearrangement phylogeny methods require that the genomic content be the same in all the organisms being compared, so that every marker (whether gene, anchor, probe binding site or chromosomal segment) in one genome be identified with a single orthologous counterpart in each of the others, though adjustments can be made for a limited amount of marker deletion, insertion and duplication.

Many genomes have been shown to result from an ancestral doubling of the genome, so that every chromosome, and hence every marker, in the entire genome is duplicated simultaneously. The present-day genome, which we refer to as a doubling descendant, can be decomposed into a set of duplicate or near-duplicate markers dispersed among the chromosomes.

There is no direct way of partitioning the markers into two sets according to which ones were together in the same half of the original doubled genome. Genomic distance or rearrangement phylogeny algorithms are not applicable to doubling descendants, since there is a two-to-one relationship between markers in the doubling descendant and related species whose divergence predates the doubling event, or unresolved two-to-two relationships between two doubling descendants, whereas these algorithms require a one-to-one correspondence.

We study rearrangement phylogeny where doubling descendants are considered along with related unduplicated genomes. We focus on the small phylogenetic problem, i.e., identifying the ancestral genomes for a given phylogeny that jointly minimize the sum of the rearrangement distances along its branches.

We review the methodology for the small phylogeny problem using gene order data, based on the iterative application of the median algorithm successively to all the ancestral vertices of a phylogenetic tree. When doubling descendants are considered together with unduplicated genomes, there are four kinds of median problem depending whether zero, one, two or three doubling descendants are among the three input genomes. We propose algorithms for handling each

case. We also analyze two ways of relating genomes from two doubling descendants, one where they result from a single genome doubling event followed by a speciation, and the other where speciation precedes two genome doublings, one in each lineage. We have applied our methods to a large data set on yeast. [2, 1, 3]

References

[1] D. Sankoff, C. Zheng, and Q. Zhu. Polyploids, genome halving and phylogeny. *Bioinformatics*, 23:i433–i439, 2006.

[2] C. Zheng, Q. Zhu, and D. Sankoff. Genome halving with an outgroup. *Evolutionary Bioinformatics*, 2:319–326, 2006.

[3] C. Zheng, Q. Zhu, and D. Sankoff. Parts of the problem of polyploids in rearrangement phylogeny. In *RECOMB 2007 Comparative Genomics Satellite. G. Tesler G, Durand D (eds) Lecture Notes in Computer Science 4751, Springer*, pages 162–176, 2007.

Reconstruction of ancestral chromosomes : methodological frameworks

Cedric Chauve[1], **Eric Tannier**[2]

[1]Simon Fraser University, Vancouver, Canada.
[2]INRIA, Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon 1, France.

We import the methodology from physical mapping and assembly of genomes to the problems of reconstruction of ancestral karyotypes. We test some algorithms on vertebrate data, and compare the results to other studies, that use minimum distance approaches.

The complexity of deriving multi-labeled trees from bipartitions

V. Moulton, K.Huber, M.Lott and A.Spillner

School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK.

Recently, multi-labeled trees have been used to help unravel the evolutionary origins of polyploid species. A multi-labeled tree is the same as a phylogenetic tree except that more than one leaf may be labeled by a single species, so that

the leaf set of a multi-labeled tree can be regarded as a multiset. In contrast to phylogenetic trees, which can be efficiently encoded in terms of bipartitions of their leaf sets, we show that it is NP-hard to decide whether a collection of bipartitions of a multiset can be represented by a multi-labeled tree. Even so, we also show that it is possible to generalize to multi-labeled trees a well-known condition that characterizes when a collection of bipartitions encodes a phylogenetic tree. Using this generalization, we obtain a fixed-parameter algorithm for the above decision problem in terms of a parameter associated to the given multiset.

Comparison of uncorrelated and autocorrelated Relaxed Phylogenetics

Michael Defoin-Platel and Alexei Drummond

Department of Computer Science, University of Auckland, Private Bag 92019, Auckland, New Zealand.

For inferring divergence time in phylogenetics, it is convenient to assume a constant molecular evolutionary rate over time. This assumption is called the molecular clock hypothesis and provides a means to translate genetic distances into geological times. Deviation from clocklike evolution has been often reported in datasets and the molecular clock hypothesis is therefore violated, particularly when distantly related species are compared, potentially leading to not only an incorrect estimation of species divergence time but also an incorrect inference of phylogenies. In the context of Bayesian phylogenetics reconstruction and divergence date estimation, it is now common to allow every branch to have a different rate of molecular evolution.

We propose to review several existing approaches to relax the molecular clock assumption, such as local clock models, auto-correlated and uncorrelated relaxed-clock models. Clock models comparisons are performed using the software Beast for numerical simulations and some particular questions are addressed such as: How good are the models and what statistics can be used for the comparison? How much different are the divergence time estimations? How the models influence the retrieved phylogenies? How the models deal with extremely long sequences?

Bayesian estimation of selection pressure on protein coding sequences

Aude Grelaud, Franois Rodolphe, Christian P. Robert

INRA MIG; Paris Dauphine University, France.

Methods for detecting Darwinian selection at the molecular level rely on estimating the rates or numbers of non synonymous changes in an alignment of protein coding sequences. The parameter of interest, called omega, represents the ratio d_N/d_S where d_N is the number of nonsynonymous changes and d_S the number of synonymous changes. The larger omega is, the strongest the evidence is that some non synonymous mutations offer fitness advantages and tend to be fixed in the population. We will speak about positive, adaptive or diversifying selection.

These methods are said "site specific" because they allow omega to vary along the sequence.

We developed in this work a model in a fully Bayesian framework. We assume that substitutions occur on the phylogentic tree according to a jump process. The instantaneous rate of change from codon i to codon j depends on omega, kappa, the transition tranversion ratio and pi, the frequency of the codons [2, 4]. We calculate the likelihood using the pruning algorithm [1]. Finally, there is one parameter of interest omega and nuisance parameters kappa, pi and the phylogenetic tree. We proposed several prior distributions for omega such as beta distribution, mixture with a point mass on some omega_0>1 and a Dirichlet process mixture model. Nuisance parameters estimation is included. Calculating joint posterior distribution involves summation over all possibles trees and, for each tree, integration over all combinaisons of parameter values. We use MCMC methods to approximate posterior probability distribution. We choose to represent the tree as a coalescent tree and take the Kingman's model as a prior. Sequences at internal nodes of the tree are considered as hidden variables and as a consequence, the likelihood calculation is easier. We use the branch-swapping algorithm of Wilson and Balding [3] which is more efficient than classical algorithms for phylogenetic tree estimation.

In these models, sites are assumed to be independent. In fact, adjacent sites are more likely to have similar values of omega. In a second model, in order to better biological realism, we introduce an hidden Markov model. We take into account the correlation between sites for the estimation of the parameter omega.

References

[1] J. Felsenstein. Evolutionary trees from DNA sequences : a maximum likelihood approach. *Journal of Molecular Evolution*, 17(1):368–376, 1981.

[2] J. Huelsenbeck, S. Jain, S. Frost, and S. Pond. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *PNAS*, 103:6263–6268, 2006.

[3] I. Wilson and D. Balding. Genealogical Inference From Microsatellite Data. *Genetics*, 150:499–510, 1998.

[4] Z. Yang, R.Nielsen, N. Goldman, and A-M. Krabbe Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155:431–449, 2000.

The dynamics of positive selection on the mammalian tree: A Bayesian inference of selection histories

Carolin Kosiol[1], Tomas Vinar[1], Rute R Da Fonseca[2], Melissa J Hubisz[3], Carlos D Bustamante[1], Rasmus Nielsen[2] and Adam Siepel[1]

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA.
[2]Human Genetics, University of Chicago University, Chicago, Illinois 60637, USA.
[3]University of Copenhagen, Copenhagen, Denmark.

We have conducted the most comprehensive examination of mammalian positively selected genes (PSGs) to date, using the six high-coverage genome assemblies now available for eutherian mammals and standard likelihood ratio tests (LRTs) to detect selection in several lineages and clades [2].

To gain further insight into the dynamics of positive selection on the mammalian tree, we devised a probabilistic model to describe the possible histories of positive selection in mammalian genes. Briefly, the probabilistic model is defined in terms of a simple switching process for the evolutionary mode (selected or non-selected) along the branches of the phylogeny. It has separate parameters for the rates of gain and loss of positive selection on each branch of the tree. The joint posterior distribution of these parameters and of all selection histories is inferred from the data by a Gibbs sampling algorithm. The inference procedure is computationally intensive, so it was applied only to the 544 genes that had been identified by one or more LRTs as showing significant evidence of positive selection. To reduce computational cost, the inference of selection histories was conditioned on the maximum likelihood estimates of the parameters of the codon models.

The inferred rates of gain and loss of positive selection were quite large, suggesting that genes tend to switch frequently between alternative modes of evolution. The posterior distributions over histories suggest that few genes have experienced positive selection specific to individual branches or clades. Instead, most genes appear to have switched between evolutionary modes multiple times. Indeed, the posterior expected number of mode switches per gene is 3.0 (1.2 gains and 1.8 losses). An expected 88% of PSGs have experienced two or more mode switches, and an expected 62% have experienced three or more mode switches. Thus, this analysis suggests that positive selection—at least for the PSGs that are identifiable by our methods—tends to occur in bursts or episodes, with

intervening intervals in which it is largely absent. Interestingly, our observations are qualitatively compatible with Gillespie's theoretical model of an episodic molecular clock [1], although our simple model does not describe the switching process in continuous time and has a resolution limited to individual branches of the tree. We also discuss the individual most likely selection histories for particularly interesting genes.

References

[1] JH. Gillespie. The molecular clock might be an episodic clock. *Proc Natl Acad Sci USA*, 81:8009–80013, 1984.

[2] Z. Yang. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*, 19:908–917, 2002.

Studying historic demographic parameters using an approximate bayesian computation

Joao Lopes, Mark Beaumont

University of Reading, School of Biological Sciences, Philip Lyle Research Building, Reading RG6 6BX, UK.

Approximate Bayesian Computation (ABC) is a recent developed Bayesian technique that can be used to extract information from DNA data. This method has been firstly introduced to Population Genetics in [5]. It relies in two major approximations: the use of a simulated step that substitutes the need for using an explicitly likelihood function; and the summarization of DNA data with a set of summary statistics. This Bayesian approach can be used to estimate several demographic historic parameters from populations using DNA data. Its main advantages are the decrease on computation time demanding and the increase on efficiency and flexibility when dealing with multiparameter models.

A particular ABC method, similar to the one used by [1], is being studied against a commonly used Markov Chain Monte Carlo (MCMC) method [3] to infer the accuracy of the previous. These approaches use DNA sequence data to extract demographic information (e.g. population sizes, time of splitting events, migration rates) within a two populations "Isolation with Migration" model. In a recent approach the developed ABC method has been tested within several different conditions (e.g. data transformation; different summary statistics sets; different number of iterations; different tolerance intervals; use of PCA).

Finally the ABC framework was applied to published data of bonobos and chimpanzees [6]. This data has been studied using several flavours of MCMC [6, 4, 2].

These studies confirm the competitiveness of the recently explored Bayesian method when compared to a standard MCMC approach. Its potential role on researches with more complex, therefore more realistic, models is emphasized.

References

[1] M. Beaumont. Joint determination of topology, divergence time, and immigration in population trees. In *Simulation, Genetics, and Human Prehistory, edited by S. MATSUMURA, P. FORSTER and C. RENFREW. McDonald Institute for Archaeological Research, Cambridge*, 2007.

[2] C. Becquet and M. Przeworski. A new method to estimate parameters of speciation models, with application to apes. *Genome Research*, 17:1505–1519, 2007.

[3] J. Hey and R. Nielsen. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. *Genetics*, 167:747–760, 2004.

[4] J. Hey and R. Nielsen. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *PNAS*, 104:2785–2790, 2007.

[5] J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol*, 16:1791–1798, 1999.

[6] Y. J. Won and J. Hey. Divergence population genetics of chimpanzees. *Mol Biol Evol*, 22:297–307, 2005.

Evolutionary pathways for sex-determining mechanisms based on X-chromosome elimination: The sciarid system

Lucas Sánchez

Centro de Investigaciones Biológicas (CSIC), Ramiro de Maeztu 9, 28040 Madrid, Spain.

Sex determination refers to the developmental program that commits the embryo to either the male or the female pathway. In most of the species, the sex of an individual is fixed at fertilisation. However, in some cases the chromosomal differences determining gender are brought about by specialised behaviour of the X chromosome during the first stages of embryonic development. The differential elimination of sex chromosomes as a mechanism of sex determination is seen in dipteran sciarid flies, in which all zygotes start with the 3X;2A

constitution: oocytes are X;A and sperm are 2X;A. This is so because meiosis is ortodox in females but highly specialised in males, showing elimination of the paternally derived genome (1st division), and the non-disjunction of the two X chromatids and elimination of one chromatid of each autosome during the 2nd division. When the zygotic nuclei reach the egg cortex, one paternal X chromosome is eliminated in the somatic cells of embryos destined to be females (2X;2A) and two are eliminated in those destined to become males (X;2A). In the formation of the X/A chromosomal signal in sciarids an imprinting process occurs in one of the parents, which determines that the chromosomes to be eliminated are of paternal origin [2]. A maternal factor controls the number of X chromosomes eliminated by the zygote [3, 1, 4]. Therefore, the formation of the primary, chromosomal signal (2X;2A versus X0;2A) determining gender in sciarids is the consequence of four processes: lethality of non-X bearing sperm, non-disjunction of maternal-derived X chromatids during spermatogenesis, elimination (controlled by a maternal factor) of X paternal-derived X chromosomes in the embryo and chromosome imprinting. This work focuses on the putative evolutionary pathways that gave rise to sciarid sex determination system from the more ancient XX/X0 system, where the primary, chromosomal signal (2X;2A versus X0;2A) is a direct consequence of the chromosomal constitution of the gametes: oocytes are X;A and sperm are 0;A.

Acknowledgements: this work was financed by grants BFU2005-03000 awarded to L. Sánchez by the D.G.I.C.Y.T., Ministerio de Educación y Ciencia, España.

References

[1] B. de Saint-phalle and W. Sullivan. Incomplete sister chromatid separation is the mechanism of programmed chromosome elimination during early Sciara coprophila embryogenesis. *Development*, 122:3775–3784, 1996.

[2] SA. Gerbi. In Germ Line Soma Differentiation. In *Results and Problems in Cell Differentiation, Hennig W. (Ed.), Springer-Verlag, Berlin*, volume 13, pages 71–104, 1986.

[3] C.W. Metz. *Amer. Nat.*, 72:485–520, 1938.

[4] L. Sánchez and ALP Perondini. Sex determination in sciarid flies: a model for the control of differential X-chromosome elimination. *J. theor. Biol*, 197:247–259, 1999.

Evolutionary process of a tetranucleotide microsatellite locus in Acipenseriforms

ZhaoJun Shao[1,2], Patrick Berrebi[3], **Eric Rivals**[4], Bin Zhu[1,5], Na Zhao[1], Sovan Lek[2] and Jianbo Chang[1,5]

[1]State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, 430072, China.
[2]Ichthyology and Ecological Modelling, UMR 5174 - Lab. EDB (Evolution Diversité Biologique), University of Paul Sabatier - CNRS, 31062 Toulouse cedex 4 - France.
[3]Institut des Sciences de l'Evolution (UMR UM2-CNRS 5554) University Montpellier II, cc 065. Place E. Bataillon 34095 Montpellier Cedex 5, France.
[4]Methods and Algorithms for Bioinformatics, L.I.R.M.M., UMR 5506 CNRS - Université de Montpellier II, 161 rue Ada, F-34392 Montpellier Cedex 5, France.
[5]Institute of Hydroecology, Ministry of Water Resources, Chinese Academy of Sciences, Wuhan, 430079, China.

The tetranucleotide microsatellite locus Spl-106 has been widely used as a molecular marker in sturgeon studies. To investigate the evolutionary process of this highly variable locus in Acipenseriforms, cross-species amplifications were performed in 130 individuals from 15 species and successful in 13 species. All PCR products were sequenced. According to the flanking sequences, a total of 94 alleles at locus Spl-106 were found in 11 out of 13 species. Twenty-three haplotypic flanking sequences were detected and four of them are dominant types present in 70 out of 94 alleles. Two of the dominant types are species-specific types, and the other two are composed of alleles from species of the Pacific and Atlantic lineages, respectively. The repeat region evolved synchronously with the flanking region. The Atlantic clade was also found in the genealogy tree of the repeat region constructed using the MS_Align method. Although the basic repeat structure was variable, several alleles were highly conserved among species and evolved independently. The evolutionary process of this locus in Acipenseriforms was reconstituted from a single repeat (TAGA)n to compound repeats (TAGA)n(TAAA)m, then to another single repeat (TAAA)n, and finally to a totally new compound repeat structure (TAAA)m(GAAA)n. Reciprocally, for the sturgeon phylogeny, our results suggest that Acipenser sturio diverged earlier than Schphirhynchus platorynchus, and infirm the Huso genus, since the two Huso species are classified within the Acipenser species of the Pacific and Atlantic lineages, respectively. Moreover, the sequence information also supports the close relationship between A. sinensis and A. dabryanus, and the relationships among A. transmontanus, A. schrenckii, and H. dauricus.

Application of Matlab in population genetics and molecular evolution

James Cai

Department of Biology, Stanford University, USA.

BACKGROUND: Matlab is a high-performance language for technical computing, integrating computation, visualization, and programming in an easy-to-use environment. It has been widely used in many areas, such as mathematics and computation, algorithm development, data acquisition, modeling, simulation, and scientific and engineering graphics. However, few functions are freely available in Matlab to perform the data analyses specifically required for population genetics and molecular evolution.

RESULTS: I have developed two Matlab toolboxes, namely PGEToolbox and MBEToolbox, aimed at filling this gap by offering efficient implementations of the most needed functions for molecular population genetists. PGEToolbox handles both DNA sequence polymorphisms and single nucleotide polymorphisms (SNPs), which include genotype and haplotype data, tests neutrality based on the coalescent theory. MBEToolbox manipulates aligned sequences, calculate evolutionary distances, estimate synonymous and nonsynonymous substitution rates, and infer phylogenetic trees. Both of them are featured by: 1) extendibility and scalability for complex and large genome-wide datasets; 2) simple yet effective graphic user interfaces and sophisticated visualization of data and results.

CONCLUSION: PGEToolbox and MBEToolbox are useful tools that can aid in the exploration, interpretation and visualization of data in population genetics and molecular evolution.
PGEToolbox is available at `http://bioinformatics.org/pgetoolbox`, and MBEToolbox is available at `http://bioinformatics.org/mbetoolbox`.

Expected time to coalescence and Fst under a skewed offspring distribution among individuals in a population

Bjarki Eldon

4100 BioLabs, Harvard University, 16 Divinity ave, Cambridge, MA 02138, USA.

We consider expected time to coalescence for two neutral genes at a single locus under skewed offspring distribution among individuals in a population subjected to subdivision and migration. The population models considered are simple mixture distributions [1]. Models of subdivision include finite number of demes as well as the many demes limit [3].

Variance of time to coalescence is also derived. Following [2] formulas for a commonly used indicator of population subdivision, Fst, are also obtained. The expected times, variances, and Fst are all shown to be functions of the parameters controlling the size and frequency of the large reproduction events. The expected time to coalescence of two ancestral lines sampled from the same subpopulation is always found to be less than the expected time for two lines sampled from different demes even under high migration rate. This shows that

highly skewed offspring distribution can maintain effects similar to those of population subdivision. The predictions of Fst can range between zero and one depending on the values of the reproduction parameters and even under high migration rate. This indicates that Fst may not be a good indicator of population subdivision for organisms with highly skewed offspring distribution.

References

[1] Eldon and Wakeley. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172:2621–33, 2006.

[2] Slatkin. Inbreeding coefficients and coalescence times. *Genet. Res.*, 58:167–75, 1991.

[3] Wakeley. Segregating sites in Wright's island model. *Theor. Popul. Biol.*, 53:166–74, 1998.

Coalescence: exact calculations of coalescence times for subdivided population

Alain Franc[1], Emmanuelle Jousselin[2]

[1]INRA, UMR Biodiversité, Gènes et Communautés, 69 route d'Arcachon, Pierroton, 33612 Cestas Cedex, France.
[2]INRA, UMR CBGP, Campus International de Baillarguet, CS 30 016, 34 988 Montferrier-sur-Lez, France.

Coalescence is a mainstream research area for inferring phylogenies from current diversity. The most dominant model in discrete non overlapping generations is Fisher-Wright neutral model. We present here an exact calculation of distribution of coalescent times between two replicators in case of non neutral choice of ancestors (and, to keep it simple, haploid, although the calculation works as well for diploid organisms). This includes the well known analytical results for island model. Exactness should not be understood here as analytical. The distribution times can be recovered from powers of some matrices, of reasonable size, modelling ancestor search as a Markov process and coalescence as an absorbing state. The calculation is exact if the calculation of the powers of the matrix is exact. It is analytical in case of the island model, with equal sizes in demes, and equal migration rates between demes. Isolation by distance is recovered, as well as current topic of isolation by resistance. Some applications will be presented, as coalescent times within metapopulations with arbitrary migration matrix between panmictic demes, not necessarily of the same size, or possibly cophylogenies between hosts and parasites. A link between the result of coalescence times and some properties of graphs describing the geometry of migration routes or host switch and their intensity will be presented.

Summarizing Multiple Gene Trees Using Cluster Networks

Regula Rupp, Daniel Huson

University of Tübingen, Germany.

The result of a multiple gene tree analysis is usually a number of different tree topologies that are each supported by a significant proportion of the genes. We introduce the concept of a cluster network that can be used to combine such trees into a single rooted network, which can be drawn either as a cladogram or phylogram. In contrast to split networks, which can grow exponentially in the size of the input, cluster networks grow only quadratically. A cluster network is easily computed using a modification of the tree-popping algorithm, which we call network-popping. The approach has been implemented as part of the Dendroscope tree-drawing program and its application is illustrated using data and results from three recent studies on large numbers of gene trees.

Level-k phylogenetic networks

Leo van Iersel, Steven Kelk, Matthias Mnich

Techinische Universiteit Eindhoven, Eindhoven, Netherlands.

Level-k phylogenetic networks are a way to describe and visualise evolutionary histories that have undergone so called reticulate evolutionary events as recombination, hybridisation or horizontal gene transfer. The level k determines how non-treelike the evolution is allowed to be. We study the problem of constructing these level-k phylogenetic networks from triplets; phylogenetic trees for triples of taxa. We give, for each k, a level-k network that is uniquely defined by its triplets. We demonstrate the usefulness of this result in a proof that the construction of level-k phylogenetic networks from triplets is NP-hard for all k. On the positive side, we give an exponential time exact algorithm for constructing level-1 networks. In addition, we show how high level networks one needs to explain any triplet set.

Phylogenetic Diversity with Disappearing Features

Charles Semple[1], Magnus Bordewich[2], Allen Rodrigo[3]

[1]Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand.
[2]Durham University, UK.

[3]University of Auckland, New Zealand.

Maximizing phylogenetic diversity (PD) is a prominent selection criteria for deciding which species to conserve or genomes to sequence. Intuitively, given a phylogenetic tree $\mathcal{T}$, the PD of a set of species is the sum of the edges of the subtree of $\mathcal{T}$ spanned by the set. Under PD, one implicitly assumes that 'features' arise at a constant rate and persist forever. But what happens if one extends this model so that features have a constant probability of disappearing? In this talk we describe some recent investigations into this question.

Encoding phylogenetic trees in terms of weighted quartets

K. Huber[1], S. Gruenewald[2], V. Moulton[1], C. Semple[3], and A. Spillner[1]

[1]School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK.
[2]CAS-MPG Partner Institute for Computational Biology (PICB), Shanghai Institutes for Biological Sciences (SIBS), Chinese Academy of Sciences (CAS), 320 Yue Yang Road, Shanghai 200031, China.
[3]Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand.

One of the main problems in phylogenetics is to develop systematic methods to reconstruct phylogenetic trees on some taxa set X. In general such trees are edge-weighted and several methods have been proposed that work by trying to piece together *quartet trees* on X, i.e. fully resolved phylogenetic trees having all four leaves in X whose interior edge is weighted. Hence, it is of interest to characterise when a collection of quartet trees corresponds to a (unique) edge-weighted phylogenetic tree. In this talk we present such characterizations that have recently been established.

Healing source trees to obtain healthy supertrees

Céline Scornavacca[1,2], Vincent Berry[1], Emmanuel J.P. Douzery[2], Vincent Ranwez[2]

[1]Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR 5506, CNRS-Université de Montpellier 2, 161 rue Ada, 34392 Montpellier Cedex 5, France.
[2]Institut des Sciences de l'Evolution de Montpellier, UMR 5554, CNRS-Université de Montpellier 2, CC 065, 34095 Montpellier Cedex 05, France.

Supertree methods combine overlapping phylogenies, rather than the primary data underlying those trees, to create a more inclusive phylogeny. When the source trees have a small overlap or a high rate of contradictions, supertree methods can propose unresolved, hence uninformative supertrees. To overcome this problem, we propose to infer non-plenary supertrees, i.e. supertrees containing only a subset of taxa, by excluding from the analysis the taxa whose position greatly differs among source trees.

We present PhySIC_IST [3], a non-plenary version of the PhySIC supertree method [1]. PhySIC returns a supertree with appealing theoretical properties: it does not propose clades contradicting the source trees (PC property) and only proposes clades present in a source tree or collectively induced by the source trees (PI property).These features provide an unambiguous phylogenetic framework that is well suited for taxonomic revisions and for applications where the reliability of the supertree is more important than its resolution.

To estimate the extent of information contained in a supertree, we also propose a new objective criterion, called the resolution degree. It takes into account both the absence of some taxa and the presence of multifurcations. PhySIC_IST is a heuristic to obtain a supertree with a maximum resolution degree while strictly satisfying the PC and PI properties.

Additionally, we propose a statistical preprocessing step (CST) to correct the source trees. This preprocess analyzes the source trees to correct prominent reconstruction artefacts due to lateral gene transfers, long branch attractions, paralogy...

Performing large-scale simulations, we observe that PhySIC_IST is, on average, at least five times more informative than PhySIC, whatever the extent of overlap among source trees, while type I error remains inferior to 1% and always inferior of that of MRP. To illustrate the effectiveness of our method on real data. We also present a biological case study centred on placental mammals. Source trees come from the OrthoMaM database [2]. PhySIC_IST with CST provide a fully resolved phylogeny agreeing with current knowledge for the group. Moreover, the CST preprocess successfully detected reconstruction artefacts in the source trees.

References

[1] V. Ranwez, V. Berry, A. Criscuolo, P. Fabre, S. Guillemot, C. Scornavacca, and E. Douzery. PhySIC: a Veto Supertree Method with Desirable Properties. *Syst. Biol.*, 56(5):798–817, 2007.

[2] V. Ranwez, F. Delsuc, S. Ranwez, K. Belkhir, M. Tilak, and E. Douzery. OrthoMaM: A database of orthologous genomic markers for placental mammal phylogenetics. In press, 2008.

[3] C. Scornavacca, V. Berry, E. Douzery, and V. Ranwez. PhySIC_IST: healing source trees to obtain healthy supertrees. Submitted BMC Bioinformatics, 2007.

Human genetic ancestry: When branches get short

Arndt Haeseler, Ingo Ebersberger, Greg Ewing, Heiko Schmidt

Center for Integrative Bioinformatics Vienna, Dr-Bohr-Gasse 9/6, A-1030 Vienna, Austria.

We discuss the complex genetic ancestry of modern humans that arose due to incomplete lineage sorting the three different gene trees are possible. We show how the frequency of each of the three trees can be used to make more sophisticated statement of our genetic relatedness to chimpanzees and gorillas. Finally, we will discuss the complexities that arise if more than three species are considered. As shown by Degnan and Rosenberg it is then possible that the most frequent gene tree is different from the species tree. We will present an approach to overcome this problem, when inferring species trees from gene trees.

The computational biology of genetically diverse assemblages

Allen Rodrigo, Helen Shearman, Frederic Berthels

Bioinformatics Institute and the Allan Wilson Centre for Molecular Ecology and Evolution, Auckland, New Zealand.

Just when we were getting used to dealing with single genomes, we now find that we have to develop tools to deal with genomic diversity. New sequencing and other bio-technologies mean that collections of molecular data from genetically diverse assemblages are fast becoming the norm. Our own work on genetically variable viruses has prompted us to look at the complexities of analysing such data. In this talk, I will discuss aspects of our ongoing work on computational metagenomic analyses. The results presented are "hot-off-the-press", but they signal the direction our research will take in the next few months/years.

Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus

E.M. Cottam[1,2], **G. Thébaud**[2,5], J. Wadsworth[1], J. Gloster[1,3], L. Mansley[4], D.J. Paton[1], D.P. King[1], D.T. Haydon[2]

[1]Institute for Animal Health, Ash Road, Pirbright, Surrey GU24 0NF, UK.

[2]Division of Environmental and Evolutionary Biology, University of Glasgow, Glasgow G12 8QQ, UK.
[3]Met Office, Fitzroy Road, Exeter EX1 3PB, UK.
[4]Animal Health Divisional Office, Strathearn House, Broxden Business Park, Perth PH1 1RZ, UK.
[5]INRA, UMR BGPI, CIRAD TA A-54/K, Campus de Baillarguet, 34398 Montpellier, France.

Estimating detailed transmission trees that reflect the relationships between infected individuals or populations during a disease outbreak often provides valuable insights into both the nature of disease transmission and the overall dynamics of the underlying epidemiological process. Genetic data are becoming increasingly important in the estimation of transmission trees, especially for viral pathogens due to their inherently high mutation rate [3].

In the case of the 2001 UK foot-and-mouth disease (FMDV) outbreak, such trees have been based on epidemiological data that relate to the timing of infection and infectiousness [1], or genetic data that show the genetic relatedness of pathogens isolated from infected individuals [2].

Here, we present a maximum-likelihood approach that allows epidemiological and genetic data to be combined within the same analysis to infer probable transmission trees. We apply this approach to data from 20 FMDV-positive farms, using complete viral genome sequences from each infected farm and information on when farms were first estimated to have developed clinical disease and when livestock on these farms were culled. Incorporating known infection links due to animal movement prior to imposition of the national movement ban results in the reduction of the number of trees from 41472 that are consistent with the genetic data to 1728, of which just 4 represent more than 95% of the total likelihood calculated using a model that accounts for the epidemiological data. These trees differ in several ways from those constructed prior to the availability of genetic data.

References

[1] D.T. Haydon et al. The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. *Proc. R. Soc. B*, 270:121–127, 2003.

[2] E.M. Cottam et al. Molecular epidemiology of the foot-and-mouth disease virus outbreak in the United Kingdom in 2001. *J. Virol.*, 80:11274–11282, 2006.

[3] R.G. Wallace et al. A statistical phylogeography of influenza A H5N1. *Proc. Natl Acad. Sci. USA*, 104:4473–4478, 2007.

Applying molecular clocks to date the intraspecific and interspecific diversification of RNA plant virus species

Denis Fargette

IRD. UMR RPB. BP 64501. Montpellier cedex 1. France.

Rice yellow mottle virus (RYMV), an RNA plant virus of the sobemovirus genus, is a major pest of rice in Africa. RYMV originated in East Africa and spread westward across the continent. Recently, the evolution rate of RYMV was estimated using heterochronous sequences of the coat protein gene (720 nt-long) of 253 isolates collected all over Africa [2]. The evolution rate was calculated under strict and uncorrelated relaxed molecular clocks as implemented in BEAST [1]. This is the first estimate of the evolution rate of an RNA plant virus. It showed that an RNA plant virus such as RYMV evolved as rapidly as most RNA animal viruses. This evolution rate of RYMV is used here to date virus events at the intraspecific and at the interspecific levels.

At the intra-specific level, the evolution rate was applied to date, under several coalescent populations models, the diversification of RYMV in Africa. It shows that the diversification started c. 200-300 years ago. This is long after rice domestication (3000 years) and rice introduction (500-1000 years) in Africa, but concomitant to rice intensification in the continent (19th century). Virus emergence in different parts of Africa was dated by downward calibration. It indicates that RYMV spread across the continent within less than one century, and that it emerged in the past few decades in West-Africa and in Madagascar.

At the interspecific level, the date of diversification of RYMV was used for upward calibration of the phylogeny of the sobemovirus genus. The sobemovirus genus comprised 10 species that have been fully sequenced (4000 to 4500 nt long). No interspecific event of recombination was detected. Intraspecific divergence reached 10% whereas interspecific divergence was between 30 and 60%. The distortion towards the hypothesis of a strict molecular clock was limited. Altogether, the sobemovirus genus is appropriate to get the first estimate of the time scale of the interspecific differentiation of plant viruses. Preliminary analyses indicate that sobemovirus diversification occurred c. 4000 years ago (+/- 1000 years). Methodological aspects and biological implications of this work will be discussed.

References

[1] A. Drummond and A. Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7:214, 2007.

[2] D. Fargette et al. Rice yellow mottle virus, an RNA plant virus, evolves as rapidly as most RNA animal viruses. In press, 2008.

A phylogenetic follow-up study of 4 individuals infected with closely related HIV-1 strains

Kristen Chalmet[a], Filip Van Wanzeele[b], Els Demecheleer[a], Kenny Dauwe[a], Jolanda Pelgrom[b], Bea Van Der Gucht[b], Dirk Vogelaers[b], Jean Plum[a], Linos Vandekerckhov[b] and Chris Verhofstede[a]

[a]AIDS Reference Laboratory and [b]AIDS Reference Center, Ghent University and Ghent University Hospital, De Pintelaan, 185, B-9000 Gent, Belgium.

Phylogenetic analysis performed using a heuristic maximum likelihood search on viral sequences obtained from 4 acutely infected individuals reveals an infection with a genetically almost identical virus. This allowed us to investigate genetic variability and disease progression in early infection with minimal interference of virus specific factors. Two of the patients were heterozygous for the 32-bp deletion in the CCR5 coreceptor gene. Both showed a slower disease progression with lower viral load levels and a reduced rate of genetic evolution compared to the patients with normal CCR5 alleles. During the 4 years of follow-up, the mean pairwise genetic distance increased with 1.20% and 1.48% in the patients with a 32-bp deletion allele compared to 2.39% and 3.13% in the patients with normal CCR5 alleles. These differences in evolutionary rates were also apparent from the corresponding maximum likelihood phylogenetic trees. The observed relation between CCR5 heterozygosity and evolutionary rate and between the viral load set point and evolutionary rate illustrates the influence of the virus replicative capacity on its genetic evolution within a host.

Comparison of commonly used methods for combining multiple phylogenetic data sets

Anne Kupczok, Heiko A. Schmidt, Arndt von Haeseler

CIBIV MFPL, Dr.-Bohrgasse 9, 1030 Wien, Austria.

Methods that combine (sequence) data for phylogeny reconstruction are classified according to their point of combination on the way from the underlying sequences to the final tree. First, superalignment methods combine the data at an early level by directly concatenating the gene alignments. Second, supertree methods reconstruct a tree for each alignment separately and combine the set of inferred trees into a so-called supertree. And third, medium level combination methods reconstruct a phylogenetic tree from intermediate results like quartets. Each of these approaches comprises different algorithms to infer the tree.

Here we present an extensive simulation study covering methods from all these three approaches. We observe that superalignment methods generally outperform the other approaches over a wide range of simulation parameters

including sparse data and gene-specific evolutionary parameters. Only if high incongruency among gene trees occurs, other combination methods show better performance than the superalignment approach. Thus, if the assumption of a tree-like evolutionary history is valid, concatenation of the data sets utilizes all available information best.

Phylogenetic Diversity with Ecological Constraints

Beáta Faller

Biomathematics Research Centre, Mathematics and Statistics Department, University of Canterbury, Christchurch, New Zealand.

A central question in conservation biology is how to measure, predict and counter the loss of biodiversity as species face extinction. A unifying idea in analysing these questions is to measure the biodiversity of a collection of species in terms of the evolutionary history that those species span in a 'tree of life'. This measure is called phylogenetic diversity (PD).

We have been working on problems that arise in analysing PD in real ecosystems. These are different random extinction models and optimization problems.

In this talk, I will summerize our findings in this field, such as the asymptotic normality of the distribution of future PD and the computational complexity of a few optimization problems.

Tools for choosing among alternative network phylogenetic inferences

Steven M. Woolley, Samuel Harrington, Alan R Templeton

One Brookings Drive Campus Box 1229 St. Louis, MO 63130, USA.

Previous work has shown that substantial differences exist among methods of inferring haplotype networks from a fixed set of input sequences [2, 1]. This leaves the end user faced with the difficult task of choosing between sometimes discrepant inferences of the phylogenetic relationships from his/her data. Few tools beyond simple visual inspection have been developed thus far, and even that is made difficult by differing file formats and software packages. We are developing a growing set of tools in order to import, explore and compare haplotype networks from a variety of methods, including: TCS, Splitstree, Union of Maximum Parsimony Trees, Neighbor-Net, Reduced Median Networks, SHRUB-GC, as well as phylogenetic tree methods. We will describe and demonstrate numerical comparisons of different networks, in order to facilitate the choice between conflicting relationships as inferred by different methods. We

will also demonstrate preliminary work toward visual comparison of alternative networks.

References

[1] I. Cassens, P. Mardulyn, and M.C. Milinkovitch. Evaluating intraspecific "network" construction methods using simulated sequence data: do existing algorithms outperform the global maximum parsimony approach? *Systematic Biology*, 54(3):363–372, 2005.

[2] S.M. Woolley, D. Posada, and K. Crandall. A Comparison of Phylogenetic Network Methods using Computer Simulation. In press, 2008.

Parallel Adaptations to High Temperatures in the Archean Eon

Bastien Boussau[1], **Samuel Blanquart**[2], Anamaria Necşulea[1], Céline Brochier[3], Nicolas Lartillot[1], Manolo Gouy[2]

[1]Université de Lyon, Université Lyon 1. CNRS. UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, Villeurbanne F-69622, France.
[2]Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR 5506, CNRS-Université de Montpellier 2, 161 rue Ada, 34392 Montpellier Cedex 5, France.
[3]Université de Provence, Aix-Marseille, Laboratory of Bacterial Chemistry, Institute of Structural Biology and Microbiology, Marseille, France.

Because fossils from the time when cellular life originated and diversified are scant and difficult to interpret, alternative means to investigate the ecology of the Last Universal Common Ancestor (LUCA) and of the ancestors of the three domains of life are of great scientific value. It was recently recognized that footprints of the effect of temperature on ancestral organisms could be uncovered in extant genomes. Accordingly, analyses of resurrected proteins predicted that the bacterial ancestor was thermophilic and that Bacteria subsequently adapted to lower temperatures. Since the archaeal ancestor is also thought to have been thermophilic, LUCA was parsimoniously inferred as thermophilic too. However, an analysis of ribosomal RNAs supported the hypothesis of a non-hyperthermophilic LUCA. Here we show that both rRNA and protein sequences analysed with advanced, realistic models of molecular evolution provide independent support for two phases in the history of environmental temperature changes over the tree of life: in the first period, thermotolerance increased from a mesophilic LUCA to thermophilic ancestors of Bacteria and of Archaea-Eukaryota; in the second period, it decreased. Therefore, the two

lineages descending from LUCA and leading to the ancestors of Bacteria and Archaea-Eukaryota convergently adapted to high temperatures, maybe in response to a climate change of the early Earth, and/or aided by the transition from an RNA genome in LUCA to organisms with more thermostable DNA genomes. This analysis unifies apparently contradictory results into a coherent depiction of the evolution of an ecological trait over the entire tree of life.

Is protein sequence evolution constant over time?

Carolin Kosiol[1] and **Nick Goldman**[2]

[1]Biological Statistics and Computational Biology, Cornell University, USA.
[2]EMBL-European Bioinformatics Institute, UK.

Over the years, there have been a few claims that evolution proceeds according to systematically different processes over different time scales. In this talk, we demonstrate that the arguments proposed make no sense logically. Using aggregated Markov models of protein sequence evolution, we show that experimental findings of process heterogeneity can be explained without recourse to time travel or intelligent design, thus restoring our confidence in probabilistic modelling of sequence evolution.

New Methods for Genealogical Network Inference based on Local Tree Topologies with a Set of SNP Sequences in Populations

Yufeng Wu

University of Connecticut, USA.

Partly due to recombination, genealogical history of a set of DNA sequences in a population usually can not be represented by a single tree. Instead, genealogy is better represented by a genealogical network, which is a compact representation of a set of correlated genealogical trees, each for a short region of genome and possibly with different topology. Inference of genealogical network for a set of DNA sequences has many potential applications, including association mapping of complex diseases.

In this paper, we present two new methods for reconstructing genealogical networks by inferring local tree topologies, which extend and improve the previous work by Song and Hein. We first show that their "tree scan" method can be converted to a probabilistic inference method based a simple probabilistic model, where algorithms for hidden Markov model can be applied. We then focus on developing a heuristic local tree inference method called RENT that is

both accurate and scalable to larger data. Through simulation, we demonstrate the usefulness of our methods by showing that the hidden Markov model-based method is comparable in terms of accuracy with the original method, and has the advantage of being able to compute the likelihood of data. We also show RENT is competitive in terms of accuracy, and can handle much larger data. [4, 3, 1, 2]

References

[1] J. Hein. Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosci*, 98:185–200, 1990.

[2] J. Hein. A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.*, 36:396–405, 1993.

[3] Y. S. Song and J. Hein. Constructing minimal ancestral recombination graphs. *J. of Comp. Biology*, 12:159–178, 2005.

[4] Y. Wu. New Methods for Genealogical Network Inference based on Local Tree Topologies with a Set of SNP Sequences in Populations. Manuscript, 2008.

Exact and efficient algorithms for the probability of a marker under incomplete lineage sorting

David Bryant[1] and Noah Rosenberg[2]

[1]University of Auckland, New Zealand.
[2]University of Michigan, USA.

Incomplete lineage sorting is known to complicate phylogenetic analysis of species radiations. Lineages from the same species can coalesce before the time of species divergence, leading to gene trees that are in conflict with the species tree. The standard models for the evolution of markers on a gene tree and for gene trees coalescing within species trees are computationally demanding since one has to integrate over all possible gene trees at each unlinked locus. We have developed algorithms that avoid this integration over gene trees by using a variant of Felsenstein's pruning algorithm for the likelihood of a phylogeny. Given a species tree (with divergence dates and population sizes) we can compute the probability of a single binary marker, exactly and efficiently. Both finite site and infinite site models of mutation are handled. Thus, if the data consist of a collection of unlinked binary markers (such as SNP data) we can compute the likelihood of the species tree directly, bypassing the need to consider the gene tree histories. These likelihoods can then be used for Bayesian or ML inference on the species tree and its parameters.