CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

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

ACI IMPBIO

Mathematics

of

Evolution

and

Phylogeny

Conference

June 17-18 & 20-21, 2005,

Institut Henri Poincaré, Paris

Programme & Abstracts

Mathematics of Evolution and Phylogeny Conference

Date: June 17-18 and 20-21, 2005

Place: Institut Henri Poincaré, Paris

Organizers

- **Olivier Gascuel**, CNRS, Montpellier, France
- **Mike Steel**, Univ. Canterbury, Christchurch, New Zealand

Webmanagers

- **Denis Bertrand** , CNRS, Montpellier, France
- **Samuel Blanquart**, CNRS, Montpellier, France

Secretaries

- **Celine Berger** , CNRS, Montpellier, France
- **Corinne Melancon** , CNRS, Montpellier, France
- **Isabelle Duc**, Institut Henri Poincaré, Paris, France
- **Sylvie Lhermitte**, Institut Henri Poincaré, Paris, France

Theme

This conference follows a similar workshop that was held in June 2003. 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 and methods becoming more realistic, powerful, and complex. The goal of the conference is to provide pedagogical presentations of the main subjects in the field, from basic principles to the cutting edge, with time for discussion and debate. There will be presentations by some of the leading experts in the field. Each speaker will survey a broad range of methods, techniques and results. Young scientists will also be selected, to give short talks or present posters.

Sponsors

CNRS **LIRMM** **Institut Henri Poincaré** **ACI IMPBio**

Programme

FRIDAY, June 17

10h-10h45:	Accueil - Café	
10h45-11h:	Bienvenue	
11h-12h15:	**Reconstructing the Ancestral Vertebrate Genome** Hugues Roest Crollius, *Ecole Normale Supérieure de Paris*	page 1
12h15-14h15:	Déjeuner	
14h15-15h30:	**Human Paleogenetics** Laurent Excoffier, *University of Bern*	page 2
15h30-16h:	Thé	
16h-17h15:	**Evolutionary Analysis of Measurably Evolving Populations** Allen Rodrigo, *University of Auckland*	page 3

SATURDAY, June 18

9h30-10h45:	**Trees of Genes within Species** Joe Felsenstein, *University of Washington*	page 4
10h45-11h15:	Café	
11h15-11h35:	**The Statistical Analysis of Spatially Clustered Genes Under the Maximum Gap Criterion** Rose Hoberman, *Carnegie Mellon University*	page 6
11h35-11h55:	**Modelling Prokaryote Gene Content** Matthew Spencer, *Dalhousie University*	page 7
11h55-12h15:	**Estimating Speciation and Extinction Rates: a Markov Chain Monte Carlo Approach** Nicolas Salamin, *University of Lausanne*	page 8
12h15-14h15:	Déjeuner	
14h15-15h30:	**Phylogenetic Invariants: Recent Progress and New Directions** Elizabeth Allman, *University of Southern Maine*	page 9
15h30-16h:	Thé	
16h-17h15:	**Phylogenomics and the Evolution of Gene Repertoires in Bacteria** Vincent Daubin, *CNRS – Université de Lyon*	page 10

MONDAY, June 20

TUESDAY, June 21

Talk Abstracts

Reconstructing the ancestral vertebrate genome

Olivier Jaillon[1], Jean-Marc Aury[1], Frédéric Brunet[2], Jean Weissenbach[1] and Hugues Roest Crollius[3]

1. Genoscope and CNRS UMR8030, 2 rue Gaston Crémieux, 91000 Evry, France
2. CNRS UMR5161, Ecole Normale Supérieure de Lyon, 46 alle d'Italie, 69364 Lyon cedex, France
3. DYOGEN group, CNRS UMR8541, Ecole Normale Supérieure, 46 rue d'Ulm, 75005 Paris, France

hrc@ens.fr

The genome sequence of the fish Tetraodon nigroviridis has recently been obtained at Genoscope [J04]. The anchoring of this sequence on the 21 chromosomes enabled us to demonstrate that approximately 300 million years ago the genome entirely duplicated in one single event. This duplication has left a number of marks in most modern fish genomes. The first is the distribution of genes that still exist in two copies in the Tetraodon genome: most copies that lie on one chromosome possess a duplicate copy on a single other chromosome, thus displaying a striking visual signature of a single duplication event that affected all chromosomes. But this distribution concerns the few genes (about 2%) that have remained in two copies since the duplication.

A second and more general signature of the duplication can however be found by comparisons with genomes from other species. We postulated that following the duplication, the ancestral teleost genome may have been transformed in a similar way to that of the yeast Sacharomyces cerevisae genome [KBL04]. In this scenario, after the duplication took place, the genome gradually returns to a non-duplicated state by deleting the majority of redundant gene copies. This process takes places in alternation between each of the "sister" chromosomes that result from the duplication. At the end of this process, the two sister chromosomes each possess about 50% of the genes from ancestral pre-duplication chromosome and as such, do not share much similarity despite their common origin. This situation can be exploited to reconstruct ancient genome organisations by comparisons to a non-duplicated genome that shares large numbers of genes of common ancestry (orthologous genes) with Tetraodon, such as the human, mouse or chicken genome. The method consists in identifying the regions of the human genome that consistently map to two regions of the Tetraodon genome (Double Conserved Synteny, DCS).

We have developed a method to automatically identify DCSs in the human and chicken genome, using a common set of 5000 orthologous genes. Strikingly, DCS blocks cover the vast majority of the human or chicken genomes, thus further supporting the occurrence of a duplication that affected the entire ancestral genome common to humans, birds and fish. By pushing the same reasoning further, DCS blocks distributed in the human genome that consistently point to the same two Tetraodon chromosomes should originate from the same ancestral chromosome, which would have been fragmented in the human lineage. By assembling DCS blocks that point to the same two Tetraodon chromosomes, we can indeed form 12 groups that should correspond to the 12 ancestral chromosomes. We further propose different models of characteristic gene distribution in the Tetraodon genome depending on different rearrangement events between chromosomes: recent and ancient fusion and fission, or simply absence of rearrangements. This shows that the Tetraodon genome has remained remarkably stable during evolution, having been submitted to only 10 major inter-chromosome rearrangements. Together with finer reconstructions of more recent ancestral genomes (e.g. mammals, [BPT04]), these results provide the basis of a general framework in which vertebrate genome evolution can be better understood.

References

[BPT04] G. Bourque, P.A. Pevzner, and G. Tesler. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res.*, 14:507–516, 2004.

[J04] O. Jaillon *et al.* Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto karyotype. *Nature*, 431:946–957, 2004.

[KBL04] M. Kellis, B.W Birren, and E.S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428:617–624, 2004.

Human Paleogenetics

Laurent Excoffier, Nicolas Ray, and Mathias Currat

University of Berne
CMPG, Zoological Institute,
Baltzerstrasse 6, 3012 Berne, Switzerland
laurent.excoffier@zoo.unibe.ch, nicolas.ray@zoo.unibe.ch, mathias.currat@zoo.unibe.ch

Neutral genetic diversity is greatly influenced by past demography, and it is thus tempting to try to infer past demographic events from observed patterns of diversity within and between populations. Humans apparently had a complex demography with a series of range expansions, contractions, and admixture events [E02], which are difficult to handle analytically.

We have therefore implemented a realistic and spatially explicit simulation scheme to model past human demography and resulting genetic diversity [CRE04]. The simulations are split into two distinct parts: the first one simulates demography (local population densities and migrations) using a forward process and the second one uses a backward coalescent approach to simulate genetic diversity based on the demography recorded during the first phase.

This approach is quite flexible and allows one to model complex scenarios, and even to incorporate environmental constraints like continental contours, deserts, forests or mountains. We shall present two examples of these simulations having led to new results concerning the origin of modern humans and their relationship with previous representatives of the *Homo* genus.

The first application will bear on the settlement of Europe by modern humans and their potential hybridization with Neanderthals [CRE04]. At odds with previous approaches assuming an instantaneous colonization of Europe and no population structure, we find that a progressive range expansion should lead to a large Neanderthal contribution to the current European gene pool, even in case of minute initial Neanderthal input. The observed current lack of Neanderthal genes in Europe is therefore only compatible with no or extremely little hybridization between the two sub-species despite their 15,000 years of cohabitation.

The second application will bear on the geographic origin of modern humans [RCBE05]. We shall examine whether the current pattern of genetic diversity among populations is affected by the exact location of the cradle of mankind, or by the mode of speciation (unique origin vs. multiregional evolution). Using a large data set of microsatellite markers in 22 worldwide populations, we define the geographic origin of modern humans as the location of a worldwide range expansion that maximizes the fit between observed and simulated data. We shall also look at the consequences of a potential ascertainment bias in the choice of genetic markers on the location of this origin.

References

[CE04] Currat M, and L Excoffier. Modern Humans Did Not Admix with Neanderthals during Their Range Expansion into Europe. *PLoS Biology* 2(12): 2264-2274, 2004

[CRE04] Currat M, Ray N, and L Excoffier. SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Molecular Ecology Notes* 4(1): 139-142, 2004

[E02] Excoffier L. Human demographic history: refining the recent African origin model. Current Opinion in Genetics and Development 12: 675-682, 2002

[RCBE05] Ray N, Currat M, Berthier P, and L Excoffier. Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. *Genome Research* (In press), 2005

Evolutionary Analysis of Measurably Evolving Populations

Allen Rodrigo

Bioinformatics Institute and The Allan Wilson Centre for Molecular Ecology and Evolution
University of Auckland
Private Bag 92019, Auckland, New Zealand
a.rodrigo@auckland.ac.nz

A population is said to evolve measurably if, when sequences are obtained over time, there is a significant accumulation of substitutions. Examples of Measurably Evolving Populations (MEPs) include rapidly evolving viruses, and populations from which it is possible to obtain ancient DNA sequences. In this presentation, I will review the methods that have been developed, to date, to study the evolutionary genetics of MEPs. In particular, I will look at (a) phylogenetic methods, including the reconstruction of serial sample phylogenies, the estimation of evolutionary rate(s) from one or more populations and the estimation of selection, and (b) coalescent methods to estimate population size, mutation and migration rates. I will conclude with a discussion of where our research is heading.

References

[DFR01] A. Drummond, R. Forsberg, and A.G. Rodrigo. Estimating stepwise changes in substitution rates using serial samples. *Molecular Biology and Evolution*, 18:1365–1371, 2001.

[DNRS02] A. Drummond, G.K. Nicholls, A.G. Rodrigo, and W. Solomon. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161:1307–1320, 2002.

[DPR+03] A. Drummond, O.G. Pybus, A. Rambaut, R. Forsberg, and A.G. Rodrigo. Measurably evolving populations. *Trends in Ecology and Evolution*, 18:481–488, 2003.

[DR00] A. Drummond and A.G. Rodrigo. Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample upgma (supgma). *Molecular Biology and Evolution*, 17:1807–1815, 2000.

[ENR04] G. Ewing, G. Nicholls, and A.G. Rodrigo. Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations (meps). *Genetics*, 168:2407–2420, 2004.

[RF99] A.G. Rodrigo and J. Felsenstein. Coalescent approaches to hiv-1 population genetics. In K.A. Crandall, editor, *The Evolution of HIV*. Johns Hopkins University Press, 1999.

[RGF+03] A.G. Rodrigo, M. Goode, R. Forsberg, H. Ross, and A. Drummond. Inferring evolutionary rates using serially sampled sequences from several populations. *Molecular Biology and Evolution*, 20:2010–2018, 2003.

[RSD+99] A.G. Rodrigo, E.G. Shpaer, E.L. Delwart, A. K. N. Iversen, M.V. Gallo, J. Jxrgen Brojatsch, M.S. Hirsch, B.D. Walker, and J.I. Mullins. Coalescent estimates of hiv-1 generation time in vivo. In *Proceedings of the National Academy of Science, USA*, volume 96, pages 2187–2191, 1999.

Trees of Genes within Species

Joe Felsenstein

Department of Genome Sciences and Department of Biology
University of Washington
Box 357730
Seattle, WA 98195-7730, USA
joe@gs.washington.edu

The models used in inferring phylogenies typically assume that individual lineages change by a mutational process, and in effect that they consist of one copy of a gene. But these lineages are actually species, often diploid species, so that this amounts to assuming that the population size $N = 1/2$. When we consider the evolution of sequences within a single random-mating population, we have to consider the tree of descent of single copies of genes. In the standard Wright-Fisher model of theoretical population genetics, each copy at a locus in effect chooses independently which copy is its ancestor in the parental generation. Going back in time, lineages combine by being replicated from the same ancestral copy. Two lineages have a probability $1/(2N)$ each generation of coming from the same copy in the preceding generation. This was known to Sewall Wright in the 1930s, but its generalization to k copies was not developed until the 1980s. The tree of ancestry of a sample of k copies is well-approximated by J.F.C. Kingman's *n-coalescent* process (usually called "the coalescent"). This generates the genealogy by going back from k copies an exponentially-distributed length of time, with mean $4N/(k(k-1))$ generations, and then combining two randomly-chosen lineages. This continues, with independent choices each time and with the value of k falling by one with each such coalescence.

This is an excellent approximation to the genealogical process as long as there is no recombination within the sequences and as long as $k^2 \ll N$. The coalescent process can be obtained as a limit of the Wright-Fisher genealogical process as $N \to \infty$, provided that we follow it on a time scale whose units are N generations.

Simple modifications of the usual coalescent allow treatment of changes in population size, and of migration. If population size at time t generations ago is $N(t)$, then the instantaneous rate of coalescence is $k(k-1)/(4N(t))$ if there are k lineages at that time, which is $N(0)/N(t)$ times as much coalescence. With migration among multiple populations, there is a coalescent process within each population, but with the addition of a constant risk of migration events for each lineage, with a rate m_{ij} of events in a given lineage in population i in which it is seen to have just arrived from population j. Going backward in time, when a lineage is traced to population j, it increases the number of lineages available to coalesce there.

When ordinary genetic recombination is allowed, the tree of genes becomes a network, generated by a collection of trees that hold at individual sites. These are summarized by an "ancestral recombination graph" which has branches forking backwards in time when recombinations occur, with an indication of which sets of sites are inherited through which of the lineages. The rate of recombination between two sites needs only be large enough that $r > 1/(4N)$ for the trees at the two sites to be substantially different. This makes it clear that the coalescent tree at any one site is not in any sense "the tree" of ancestry of the human species. Instead there are hundreds of thousands of substantially different trees for different parts of the genome. Thus there may be a Mitochondrial Eve, but there is also a Y-chromosome Adam (who did not know Eve), a Hemoglobin Harriet, a Cytochrome Sam, and many others as well. Models involving recombination provide a sound parametric approach to within-species networks of haplotypes. They enable us to choose sound and statistically efficient methods of inference when we might otherwise have been tempted to consider arbitrary methods that allow for some loops in the network.

When we consider population samples taken from related species, there will be a coalescent process within each species, back to the time when we reach its common ancestor with another species. At that time there may be m copies ancestral to our sample from species 1, and n copies ancestral to our sample from species 2. Then we have $m + n$ samples from the common ancestor population, and as we go back from there, these coalesce. As it takes about $4N$ generations on average for a sample in a single population to completely coalesce, if the branch lengths in the tree are much larger than $4N$ generations, the coalescent trees will each be consistent with the species tree. If some of the branch lengths are smaller than $4N$ generations, it becomes more likely that there will be discrepancies between the gene tree and the species tree. In organisms such as bacteria or protists, the vary large population sizes make it at least possible that some of the discrepancies we see between trees from different loci are coalescent effects rather than horizontal gene transfer between species.

With one region of less than 1000 bases evolving according to neutral mutation, the number of segregating sites is likely to be small in any population sample. The coalescent tree will then be poorly estimated. If we knew the coalescent tree precisely, this would yield simple estimates of quantities such as the scaled population size $4N\mu$, the scaled population growth rate g/μ, the scaled migration rates $4Nm_{ij}$, and the scaled recombination rate per base, $4Nr$. Not knowing the tree precisely, we instead must compute likelihoods (or do Bayesian inference) by making the appropriate weighted average over all possible coalescent trees. When trees have branch lengths that are given in units of expected mutations per site, the likelihood for parameter values β is the integral over all tree shapes and all possible branch lengths:

$$L \;=\; \text{Prob}\,(D\,|\,\beta) \;=\; \sum_G \text{Prob}\,(G\,|\,\beta)\,\text{Prob}\,(D\,|\,G)$$

where the summation is summation over tree shapes (labelled histories) and integration over lengths of coalescent intervals for each shape. For samples of more than two sequences, there is no closed-form formula for the likelihood. It is necessary to sum over all labelled histories, and for each to integrate over all combinations of coalescence times. For 10 sequences, there are 2.571×10^9 labelled histories, and for each we need to compute a 9-dimensional integral. Although we could imagine Monte Carlo integration by sampling from the Kingman prior density of trees, this is wildly inefficient since most such gene trees have very small values of $\text{Prob}\,(D\,|\,G)$.

A better approach is importance sampling by Markov Chain Monte Carlo integration. The first such sampling method was that of Griffiths and Tavaré [GT94], who used an original independent sampling method. Our own method [KYF95] wanders through tree space, with trees proposed by small rearrangements. These have been followed by a number of different approaches, some Bayesian. Among the cases that have been treated are coalescents with population growth, migration, recombination, and species trees.

When there are L loci, all far apart enough in the genome to have independent coalescent trees, the overall likelihood of a species tree S is the product across loci of the sums over all possible genealogies G at that locus:

$$L \;=\; \text{Prob}\,(D\,|\,S,\beta) \;=\; \prod_{i=1}^{L}\Big(\sum_G \text{Prob}\,(G\,|\,S,\beta)\,\text{Prob}\,(D^{(i)}\,|\,G)\Big)$$

where $D^{(i)}$ is the data for the ith locus. This likelihood calculation (or a closely analogous Bayesian calculation) will be more statistically efficient than trying to infer species trees from gene trees by more *ad hoc* approaches such as consensus.

With this new generation of methods, we can infer, not just point estimates of parameters, but likelihood curves for them or Bayesian posteriors for them. Further explanations of the logic of coalescents can be found in the recent book by Hein, Schierup, and Wiuf [HSW04] and the forthcoming book by Wakeley [Wak05]. My own book on phylogenies covers coalescents and their inference methods [Fel04].

References

[Fel04] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts, 2004.

[GT94] R. C. Griffiths and S. Tavaré. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London, Series B*, 344(1310):403–410, 1994.

[HSW04] J. Hein, M. H. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution. A Primer in Coalescent Theory*. Oxford University Press, Oxford, 2004.

[KYF95] M. K. Kuhner, J. Yamato, and J. Felsenstein. Estimating effective population size and mutation rate from sequence data using metropolis-hastings sampling. *Genetics*, 140(4):1421–1430, 1995.

[Wak05] J. Wakeley. *Coalescent Theory: An Introduction*. Roberts and Company, Greenwood Village, Colorado, 2005.

The Statistical Analysis of Spatially Clustered Genes Under the Maximum Gap Criterion

Rose Hoberman*, David Sankoff and Dannie Durand

*Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
roseh@cs.cmu.edu, sankoff@uottawa.ca, and durand@cs.cmu.edu

Identifying gene clusters, genomic regions that share local similarities in gene organization, is an essential prerequisite for many genomic analyses, including reconstruction of chromosomal rearrangements, phylogenetic inference, and detection of whole-genome duplications, as well as identification of operons, horizontal transfer, and functional selection in bacteria.

Our goal is to provide formal statistical models to test the hypothesis that two genomic regions in distantly related genomes share a common ancestor, against a null hypothesis of random gene order. In diverged genomes, such regions are characterized by similar gene content, but neither content nor order are strictly preserved. We refer to these regions as "gene clusters." A number of formal definitions of gene clusters have been proposed [SVdP04], as well as methods for finding such clusters and/or statistical tests for determining their significance [DS03]. Unfortunately, there is very little overlap between cluster definitions used in analyses of genomic data and the definitions upon which rigorous analytical statistical tests are based.

We consider the max-gap cluster, a definition that has emerged as perhaps the most popular in empirical studies. In a max-gap cluster the distance between genes in the cluster is constrained to be no more than a constant g. Although this is one of the models most widely used in practice, and efficient algorithms have been developed to search for such clusters [BCR02], no formal statistical model has yet been developed for this cluster definition.

We present the first formal, rigorous statistical treatment of max-gap clusters. Models are developed for two basic cluster finding scenarios. In the first scenario, we wish to find clusters of a subset of m pre-specified genes (the "marked" genes), which may be of interest because their homologs are contiguous in another region or genome or because they share some functional properties. In the second scenario, the set of genes in a cluster emerges from the comparison of two whole genomes; we are given two genomes and a mapping between their homologs, and we wish to find clusters of homologs found in close proximity in both genomes.

For the marked genes scenario, we present an exact expression for the probability of observing a complete max-gap cluster containing all m marked genes within a randomly ordered genome of size n. Next we extend this analysis to evaluate the probability of observing a cluster containing only a subset of the marked genes. We present an algorithm that calculates the exact probability of observing an incomplete cluster of size $h < m$, as well as an analytic solution for the case where $m/2 < h$. In addition, we develop an upper bound for the probability of observing a max-gap cluster of h homologs by pairwise whole genome comparison scenario. Finally, we show that the precise formalization of the max-gap cluster definition has surprising implications both for designing algorithms to find clusters and testing cluster significance.

To investigate trends in the cluster probabilities for both models, we use the derived equations to calculate the probability of clusters for common choices of parameter values, and for a range of typical sizes of prokaryotic and eukaryotic genomes. We discuss the influence of genome size, gap size, the number of marked genes and the cluster size on cluster significance, and determine the regions of the parameter space that yield clusters that are statistically significant.

References

[BCR02] A. Bergeron, S. Corteel, and M. Raffinot. The algorithmic of gene teams. In D. Gusfield and R. Guigo, editors, *WABI*, volume 2452 of *Lecture Notes in Computer Science*, 2002.

[DS03] D. Durand and D. Sankoff. Tests for gene clustering. *J Comp Biol*, 10(3-4):453–82, 2003.

[SVdP04] Cedric Simillion, Klaas Vandepoele, and Yves Van de Peer. Recent developments in computational approaches for uncovering genomic homology. *Bioessays*, 26(11):1225–35, Nov 2004.

Modelling prokaryote gene content

Matthew Spencer, Edward Susko, and Andrew J. Roger

Department of Mathematics and Statistics & Department of Biochemistry and Molecular Biology,
Dalhousie University, Halifax, Nova Scotia, Canada
matts@mathstat.dal.ca, susko@chase.mathstat.dal.ca, aroger@dal.ca

The patchy distribution of genes across the prokaryote phylogeny may be due either to multiple gene losses or to lateral transfer. Many existing phylogenetic methods for gene content data use either parsimony or ad-hoc distance methods [SBH99], and are therefore uninformative about the relative rates of loss and transfer. Explicit models for gene gain and loss are necessary if we want to understand the processes that determine genome size.

The number of genes in a family in a genome can increase by duplication, lateral transfer, or innovation (evolution of a new member of a gene family from some other sequence), and can decrease by deletion. As far as we know, all existing likelihood-based methods for gene content data use birth-death models, in which only one gene can be gained or lost at a time [GZ04, ZG04, ABLS03]. There is extensive empirical evidence that duplications, deletions and lateral transfers can involve more than one gene at a time. We therefore develop models that allow multi-gene events. For simplicity, all our models are formulated as continuous-time Markov chains, with the simplifying assumptions that gene families are independent and that the maximum possible number of members of a family in a genome is finite.

Using likelihood ratio tests, we show that models allowing multi-gene events are significantly better than birth-death models, for two pairs of genomes from the COG (Clusters of Orthologous Gene families) database [TFJ$^+$03] that we studied in detail: two closely related *E. coli* strains, and *Bacillus subtilis* (bacteria) and *Archaeoglobus fulgidus* (archaea). For both pairs, the estimated rate of lateral transfers of more genes than could be gained by duplication or innovation is not significantly greater than zero. This does not imply that lateral transfers do not occur, but it does suggest that such events may only rarely transfer many genes from the same family.

We use the model that allows multi-gene events to estimate the residence time (from appearance by duplication, innovation or transfer to deletion) of a gene in a genome. Our results suggest that a substantial proportion of genes will have residence times that are too short to allow deep phylogenetic reconstruction from single-gene sequence data.

We fit both models to all pairs of genomes in the COG database, and use the resulting estimates of pairwise evolutionary distances to reconstruct least-squares phylogenies. The phylogeny based on the birth-death model does not have each of the three kingdoms (eukaryotes, bacteria, and archaea) as a monophyletic clade. The phylogeny based on a model that allowed multi-gene events is more biologically plausible. Nevertheless, it has some obvious problems, such as a tendency to group parasites and endosymbionts together. This also occurs in other gene content phylogenies [GZ04, WRGK02], and is probably due to parallel gene loss in these species.

References

[ABLS03] L. Arvestad, A.-C. Berglund, J. Lagergren, and B. Sennblad. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics*, 19 Suppl. 1:i7–i15, 2003.

[GZ04] X. Gu and H. Zhang. Genome phylogenetic analysis based on extended gene contents. *Molecular Biology and Evolution*, 21(7):1401–1408, 2004.

[SBH99] B. Snel, P. Bork, and M. A. Huynen. Genome phylogeny based on gene content. *Nature Genetics*, 21:108–110, 1999.

[TFJ$^+$03] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, 2003.

[WRGK02] Y. I. Wolf, I. B. Rogozin, N. V. Grishin, and E. V. Koonin. Genome trees and the Tree of Life. *Trends in Genetics*, 18(9):472–479, 2002.

[ZG04] H. Zhang and X. Gu. Maximum likelihood for genome phylogeny on gene content. *Statistical applications in genetics and molecular biology*, 3(1):article 31, 2004.

Estimating speciation and extinction rates: a Markov chain Monte Carlo approach

Nicolas Salamin

Department of Ecology and Evolution
University of Lausanne
1015 Lausanne - SWITZERLAND
nicolas.salamin@unil.ch

Phylogenetic trees are important tools to investigate diverse aspects of evolutionary biology, from gene evolution to factors affecting species diversification. From a macro-evolutionary point of view, the phylogenetic relationships can give information on the mode of lineage evolution, through the use of comparative methods for example. On the other hand, insights into the tempo of lineage evolution can be gained by using the temporal dimension that is associated with the branch lengths defining each tree based on molecular data. In particular, the distribution of node intervals contains all the required information to estimate the rates of speciation and extinction of species.

A stochastic process, the pure birth and death process, is used to model the tempo of species evolution through time. If each new lineage appearing in the tree is considered independent, it becomes possible to estimate the probability that this lineage does not speciate, nor get extinct over a time period that is equal to the time duration from the split event to the present time [Tho75]. A maximum likelihood approach can then be used to estimate the parameters of the birth and death process over all lineages present in the tree [HMN94, Fel04].

If the uncertainties inherent to phylogenetic reconstruction are ignored, and a single tree is used to evaluate the macro-evolutionary parameters, a potentially important bias can be introduced in the estimation procedure. A way to factor out these uncertainties while estimating the rates of speciation and extinction is to base the estimation on a set of trees that are probable for the data. The estimation of the parameters can then simply be approximated as an average over all these trees.

To efficiently sample trees, Markov chain Monte Carlo (MCMC) is used, and the most probable trees for the data are selected according to their posterior probability. MCMC application is best known through the Bayesian approach of tree reconstruction. However, to avoid the use of any prior distributions, an importance sampling scheme is developed here [GRS96]. Driving values for the parameters of the speciation model are used to evaluate the trees sampled during the MCMC, and the tree sampling strategy is adjusted through a careful calculation of the Hasting ratio. The ideal driving values should be the, unknown, maximum likelihood estimates for the parameters. To overcome this problem, regular updates of the driving values are performed through the run of successive chains.

Computer simulations were performed to test the proposed method. Three different sets of speciation and extinction parameters were used to simulate random trees under the birth and death process. DNA sequences were then simulated on these trees, and the resulting matrices analysed with the method described above. The results show that accurate estimates of speciation and extinction rates can be obtained using this method. However, large DNA matrices and comprehensive sets of terminal taxa have to be sampled in order to achieve a good accuracy. Finally, further refinements to the method are proposed, and future applications are discussed.

References

[Fel04] J. Felsenstein. *Inferring phylogenies*. Sinauer Associates, Sunderland, MA, USA, 2004.

[GRS96] W. R. Gilks, S. Richardson, and V. Spiegelhalter. *Makov chain Monte Carlo in practice*. Chapman and Hall, London, UK, 1996.

[HMN94] P. H. Harvey, R. M. May, and S. Nee. Phylogenies without fossils. *Evolution*, 48(3):523–529, 1994.

[Tho75] E. A. Thompson. *Human evolutionary trees*. Cambridge University Press, Cambridge, UK, 1975.

Phylogenetic Invariants: Recent Progress and New Directions

Elizabeth S. Allman and John A. Rhodes

Department of Mathematics and Statistics
University of Southern Maine
96 Falmouth Street
Portland, ME 04104 – USA
eallman@maine.edu

Department of Mathematics
Bates College
Andrews Road
Lewiston, ME 04104 – USA
jrhodes@bates.edu

Using polynomials in the expected pattern frequencies in aligned sequence data for phylogenetic inference was first proposed in 1987 ([CF87], [Lak87]). For a 2-state model of base substitutions and four taxa, Cavender and Felsenstein established for each of the three quartet trees the existence of quadratic polynomials that vanish when evaluated on pattern frequencies arising from the correct tree. Such polynomials are known commonly as 'phylogenetic invariants'.

The idea underlying Cavender and Felsenstein's paper was a geometric one: Assuming a particular model of molecular evolution along a tree, the expected pattern frequencies are polynomial functions in the stochastic parameters (the entries of the root distribution vector and the transition matrices associated to each edge) and thus satisfy certain polynomial relationships. In the language of algebraic geometry, the map from stochastic parameter space to the space of pattern frequencies is a parameterization that defines a high-dimensional surface, an affine algebraic variety, and testing for the vanishing of phylogenetic invariants determines whether we have located a point on this variety.

Although providing the possibility of a new viewpoint to the problem of inference, phylogenetic invariants developed slowly. It was difficult to identify invariants for models of mutation that incorporated more biological realism (rate variation), and early simulation studies using only linear invariants showed that exceptionally long sequences were necessary for accurate inference. Nonetheless, several researchers developed methods to find invariants for a variety of models. Of particular note was the work initiated by Hendy and Penny who used Hadmard conjugation on group-based models and Ferretti and Sankoff who introduced numerical approaches to finding invariants.

This talk begins with an overview and survey of results on phylogenetic invariants, beginning with Cavender and Felsenstein's early paper and then developing a more geometric framework for understanding phylogenetic invariants. Various approaches to identifying invariants will be touched upon, to highlight the array of techniques that have been used.

Next, progress in the last few years towards finding invariants for the general Markov model and group-based models will be discussed. The emphasis here will be both on explaining the results and explaining the viewpoints that led to these results. One of the main theorems for the general Markov model will emphasize that *local features* of the tree determine the phylogenetic invariants. This suggests it may be possible to develop tests with invariants to give support for bi-partitions of the taxa (splits in a tree) or tri-partitions (nodes in a trivalent tree), etc.

The final part of this talk will highlight new insights brought forth by the study of phylogenetic invariants and suggest areas for further research. In particular, a proof of the statistical identifiability of parameters for the covarion (and other) models will be mentioned, as will the potential uses of invariants in maximum likelihood methods and parameter estimation. A more complete bibliography will be available with the talk.

References

[AR03] E. S. Allman and J. A. Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Math. Biosci.*, 186:113–144, 2003.

[AR04] E. S. Allman and J. A. Rhodes. Phylogenetic ideals and varieties for the general Markov model. 2004. preprint, arXiv.org/math.AG/0410604.

[CF87] J. A. Cavender and J. Felsenstein. Invariants of phylogenies in a simple case with discrete states. *J. of Class.*, 4:57–71, 1987.

[Lak87] J.A. Lake. A rate independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Mol. Bio. Evol.*, 4:167–191, 1987.

Phylogenomics and the Evolution of Gene Repertoires in Bacteria

Vincent Daubin

Bioinformatique et Génomique Evolutive
Laboratoire BBE, UMR 5558
43 Bld du 11 Novembre, 69622 Villeurbanne cedex - FRANCE
daubin@biomserv.univ-lyon1.fr

Because phylogeneticists are in a constant quest for more phylogenetic characters, the approaches aimed at building trees using multiple genes and complete genome sequences have become very popular. Genome sequences can be used in different ways to generate pertinent phylogenetic information. The methods proposed can be classified into six groups: 1) concatenation of several sequence alignments in order to increase the statistical power of phylogenetic methods [BDI+01]; 2) clustering based on shared presence of genes, gene families or protein domains in genomes [SBH99]; 3) comparison of various statistics derived from complete genome (nucleotide composition, oligonucleotide frequencies etc...) [VA03]; 4) global indexes of genome similarity (BLAST...) [CBRC02]; 5) combination of gene trees (supertrees) [DGP02] and 6) gene order [WRG+01].

While the benefits of such methods are evident, their application supposes that certain hypotheses are respected. In prokaryotes, for instance, it is quite certain that the first of these hypotheses – that genes follow a pattern of strict vertical inheritance and thus provide reliable markers of species phylogeny – has been repeatedly violated. Intertwined with limitations due to such biological phenomenon are a number of methodological and theoretical pitfalls that need to be identified.

In this lecture, I will review recent advances in this field and show that, although the history of bacteria is not yet resolved, the phylogenomic approach has significantly highlighted the process of prokaryotic genome evolution. While the success of these methods in revealing non trivial phylogenetic relationships among bacteria is variable, the phylogenies inferred show a remarkable congruence, which contradicts recent claims that bacterial evolution would be more faithfully represented as a net than a tree. Taken together, these results have significantly clarified the respective roles of lateral gene transfer and vertical descent in bacterial evolution.

References

[BDI+01] J.R. Brown, C.J. Douady, M.J. Italia, M.E. Marshall, and M.J. Stanhope. Universal trees based on large combined protein sequence data sets. *Nat. Genet.*, 28(3):281–285, 2001.

[CBRC02] G.D. Clarke, R.G. Beiko, M.A. Ragan, and R.L. Charlebois. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized blastp scores. *J. Bacteriol.*, 184(8):2072–2080, 2002.

[DGP02] V. Daubin, M. Gouy, and G. Perriere. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.*, 12(7):1080–1090, 2002.

[SBH99] B. Snel, P. Bork, and M.A. Huynen. Genome phylogeny based on gene content. *Nat. Genet.*, 21(1):108–110, 1999.

[VA03] S. Vinga and J. Almeida. Alignment-free sequence comparison-a review. *Bioinformatics.*, 19(4):513–523, 2003.

[WRG+01] Y.I. Wolf, I.B. Rogozin, N.V. Grishin, R.L. Tatusov, and E.V. Koonin. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.*, 1(1):8, 2001.

Phylogenetic diversity: from combinatorics to conservation

Mike Steel

Biomathematics Research Centre
University of Canterbury
Private Bag 4800 - Christchurch - NEW ZEALAND
m.steel@math.canterbury.ac.nz

The 'phylogenetic diversity' (PD) of a subset of taxa in a tree is the sum of the branch lengths connecting those taxa, and it is a useful comparative measure in biodiversity conservation ([B02, F92]). This talk outlines some combinatorial and stochastic properties of PD that allow optimal PD sets to be found quickly ([NM97, S05]) and the expected loss of PD under random extinction ([MHC05, NM97]) to be quantified ([S05]). Related combinatorial properties of PD are also discussed ([PS04, SS04]). In the final part of the talk I also describe some new results concerning an extension of PD to abelian groups (joint work with Andreas Dress [DS05]), and speculate as to how this might be relevant to tree reconstruction from sequences.

References

[B02] Barker, G. M. 2002. Phylogenetic diversity: a quantitative framework for measurement of priority and achievement in biodiversity conservation. *Biol. J. Linnean Soc.* 76:165–194.

[DS05] Dress, A. and Steel, M. 2005. The path structure of X–trees (in preparation).

[F92] Faith, D.P., 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation* 61, 1–10.

[MHC05] Mooers, A.O., Heard, S.B., and Chrostowski, E. 2005. Evolutionary heritage as a metric for conservation. pp 120-138 in *Phylogeny and Conservation* (A. Purvis, T.L. Brooks and J.L. Gittleman, eds.) Cambridge University Press, Cambridge.

[NM97] Nee, S. and May, R. M., 1997. Extinction and the loss of evolutionary history. *Science*, 278: 692–694.

[PS04] Pachter, L., and D. Speyer. 2004. Reconstructing trees from subtree weights. *Appl. Math. Lett.* 17(6):615–621.

[SS04] Semple, C. and Steel, M. 2004. Cyclic permutations and evolutionary trees. *Adv. Appl. Math.* **32(4)**, 669–680.

[S05] Steel, M. 2005. Phylogenetic diversity and the greedy algorithm. *Syst. Biol.*, in press.

[S05] Steel, M. 2005. Tools to construct and study big trees: a mathematical perspective.In *Towards the tree of life: taxonomy and systematics of large and species rich taxa*, (eds. T. Hodkinson, J. Parnell and S. Waldren). CRC press (in prep).

Codon Models

Carolin Kosiol and Nick Goldman

EMBL -EBI, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK
kosiol@ebi.ac.uk, goldman@ebi.ac.uk

In the past, two kinds of Markov models have been considered to describe protein evolution. Mechanistic models are formulated on the codon level and take into account features such as transition-transversion bias, codon frequency bias and synonymous-nonsynonymous amino acid substitution bias ([GY94, YNGP00]). Empirical models do not explicitly consider factors that shape protein evolution, but attempt to summarise the substitution patterns observed from large quantities of data. To date, empirical models have only been formulated on the amino acid level, not on the codon level.

The empirical amino acid replacement models (e.g. PAM, JTT and WAG: [DSO78, KG05]; [JTT92]; [WG01]) have been widely used in protein sequence alignment, phylogeny and database searches over the last 30 years. However, amino acid sequence evolution has been shown to behave in a non-Markovian manner ([BCG94, MD95]) and non-evolutionary models such as BLOSUM ([HH95]) have been proposed. We have shown that some of the non-Markovian behaviour described in the literature can be explained by an aggregated Markov process ([Lar98]), which combines DNA mutational biases, rate heterogeneity among different codon sites of the protein, the properties of the amino acids encoded by the sequence and effects of selection operating on those amino acids. This result leads us to suggest that protein evolution should be modelled with codon-level rather than amino acid substitution models.

We have therefore decided to estimate a first empirical codon model using data taken from Pandit database ([WdBG98]) and the estimation program Dart ([HR02]). Preliminary results indicate that the evolutionary process consists of singlet, doublet and triplet changes as suggested by the 'SDT' mechanistic model ([WG04]). A grouping of the 61 sense codons into subsets with high probability of change amongst codons of each group but small probability of change between groups ([KGB04]) shows that the affiliation between a triplet of DNA and the amino acid it encodes is a main factor driving the process of codon evolution. We plan further analysis of empirical codon models, and an assessment of their performance by maximum likelihood (ML) comparison.

References

[BCG94] S. Benner, M.H. Cohen, and G.H. Gonnet. *Protein Eng.*, 7:1323–1332, 1994.

[DSO78] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. *Atlas of Protein Sequence and Structure Vol. 5, suppl. 3*, pages 345–352. National Biomedical Research Foundation, Washington, D.C., 1978.

[GY94] N. Goldman and Z. Yang. *Mol. Biol. Evol.*, 11:725–736, 1994.

[HH95] S. Henikoff and J.G. Henikoff. *PNAS*, 89:10915–10919, 1995.

[HR02] I. Holmes and G.M. Rubin. *J. Mol. Biol.*, 317:753–764, 2002.

[JTT92] D.T. Jones, W.R. Taylor, and J.M. Thornton. *CABIOS*, 8:275–282, 1992.

[KG05] C. Kosiol and N. Goldman. *Mol. Biol. Evol.*, 22:193–199, 2005.

[KGB04] C. Kosiol, N. Goldman, and N.H. Buttimore. *J. Theor. Biol.*, 228:97–106, 2004.

[Lar98] B. Larget. *J. Appl. Prob.*, 32:313–324, 1998.

[MD95] G. Mitchison and R. Durbin. *J. Mol. Evol.*, 41:1139–1151, 1995.

[WdBG98] S. Whelan, P.I.W. de Bakker, and N. Goldman. *Bioinformatics*, 19:1556–1563, 1998.

[WG01] S. Whelan and N. Goldman. *Mol. Biol. Evol.*, 18:691–699, 2001.

[WG04] S. Whelan and N. Goldman. *Genetics*, 167:2027–2043, 2004.

[YNGP00] Z. Yang, R. Nielsen, N. Goldman, and A.-M.K. Pedersen. *Genetics*, 155:431–449, 2000.

Correlation between composition and site-specific evolutionary rate and implications for phylogenetic inference

Vivek Gowri-Shankar and Magnus Rattray

School of Computer Science, Manchester University
Kilburn Building, Oxford road - Manchester - UK
vivek.gowri-shankar@cs.man.ac.uk, magnus@cs.man.ac.uk

Likelihood-based phylogenetic methods are designed around an explicit model of the sequence evolution process. Traditionally, a parametric Markov model is used to describe the nucleotide replacement process and conventional method assume that this process is the same at all sites[Fel04]. Although the presence of slow and fast evolving sites is usually accommodated by modelling the spatial variation of the evolutionary rate[Yan96], most substitution models do not account for possible spatial variation of the nucleotide equilibrium frequencies. By contrasting base pair frequencies observed at slow and fast evolving sites in the helices of contemporary RNA genes, we show that this assumption of spatial compositional homogeneity can be clearly violated. Using primates mitochondrial rRNA genes, we recover intuitive trends where G:C pairs are less frequent than A:U pairs and mismatches at fast evolving sites.

Using synthetic datasets, we explore the effects of the spatial compositional variation on evolutionary models that do not account for it. In Maximum-Likelihood and Bayesian inference, equilibrium frequencies are found to be biased toward the composition of fast evolving sites when they are estimated during the inference process. A new method is proposed to account for the compositional variation across sites. The mechanisms responsible for the variation are not modeled directly with a deterministic model. Instead a Gaussian process prior[Mac98] is used to allow for a smooth change in composition with evolutionary rate. Results suggest that this model can accurately capture the observed variations in RNA sequences.

By a daring use of statistical phylogenetic methods, Galtier et al.[GTG99] argued against a hyperthermophilic origin of life. Using a time-heterogeneous nucleotide substitution model that allows for varying G+C content over evolutionary time[GG98], they inferred the ancestral G+C composition of the rRNA genes of the last universal ancestor of extant life forms (LUCA). In contrast to the G+C content observed in contemporary thermophiles, the ancestral content appeared incompatible with a hot living environment. We investigated the behaviour of their time heterogeneous method in presence of spatial compositional heterogeneity. Synthetic sequences were generated with a stationary substitution model assuming a gamma distribution of the evolutionary rate across sites and using a different frequency vectors for each rate category. We found that the ancestral G+C content recovered with the time-heterogeneous method is biased toward the G+C content observed at slow evolving sites. The problem is exacerbated when the time-heterogeneous model is combined with a rate-across-sites variation method.

We reexamined their results in light of these findings. The G+C frequency they found was influenced by the frequency of slow evolving sites and, with the sequences they used (loops+stems of rRNA genes), the ancestral G+C content and the environmental temperature of the LUCA were probably underestimated, placing their conclusion in question. We developed a parameterised space-time heterogeneous version of the evolutionary model mentioned above. The model accounts for substitution process heterogeneity across sequences and over evolutionary time. Equilibrium frequencies used at each branch of the tree are selected among a pool of composition parameters that are assumed to vary smoothly with substitution rate. The model is tested against rRNA sequences spanning the entire tree of life. The inferred ancestral composition found does not allow to conclude either for or against a thermophilic LUCA.

References

[Fel04] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts, 2004.

[GG98] N. Galtier and M. Gouy. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Molecular Biology and Evolution*, 15:871–879, 1998.

[GTG99] N. Galtier, N. Tourasse, and M. Gouy. A nonhyperthermophilic common ancestor to extant life forms. *Science*, 283:220–221, 1999.

[Mac98] D. J. C. Mackay. Introduction to gaussian processes. In *Neural Networks and Machine Learning*, pages 133–165. Bishop, C. M., Springer, Berlin, 1998.

[Yan96] Z. Yang. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution*, 11:367–372, 1996.

The timing of eukaryotic evolution: from rocks to relaxed molecular clocks, and reciprocally.

Emmanuel J. P. Douzery, Frédéric Delsuc & Hervé Philippe

Phylogenie Moléculaire
ISE-M (UMR 5554 CNRS)
Université Montpellier II - Place E. Bataillon
34 095 Montpellier Cedex 5 - France
douzery@isem.univ-montp2.fr

Canadian Institute for Advanced Research
Département de Biochimie, Centre Robert-Cedergren,
Université de Montréal,
Montréal, Québec, Canada
frederic.delsuc@umontreal.ca
herve.philippe@umontreal.ca

The use of nucleotide and amino acid sequences to infer phylogenies allows an improved understanding of the timing of evolutionary events of life on earth. Yet, phylogenetic trees only provide a relative chronology of speciation events. The molecular clock approach circumvents this limitation by coupling molecular data with paleontological information in order to yield absolute ages of divergence between taxa. Molecular estimates of divergence times are however controversial, as they generally are either much more ancient than suggested by the fossil record, or highly variable among studies. The limited number of genes, proteins, and species explored, the pervasive variations in evolutionary rates, and the choice of geological calibrations are among the most likely sources of such discrepancies. Here we compared concatenated amino acid sequences of 129 different proteins from 36 eukaryotes to determine divergence times of several major groups, including animals, fungi, plants, and various protests [DSB04]. Due to significant evolutionary rate variations among eukaryotes, we used a relaxed molecular clock approach with rate autocorrelation along branches developed in a Bayesian framework [TKP98, KTB01]. Six simultaneous fossil constraints were incorporated as time intervals to calibrate the clock.

We showed that, according to 95% credibility intervals, some eukaryotic kingdoms diversified 950-1,259 (mean ~1,100) million years ago (Mya). Plantae (green plants + red algae) originated 892-1,182 (~1,000) Mya, animals diverged from choanoflagellates 761-957 (~850) Mya, and the debated age of the split between protostomes versus deuterostomes was estimated between 642-761 (~700) Mya [DSB04]. Bayesian divergence times and associated credibility intervals appeared robust to missing character states in the alignment and prior assumptions on the model. Interestingly, these Bayesian datings appeared to be much more recent than those obtained under the assumption of a global molecular clock, but animal diversification remained 100 million years more ancient than suggested by the Cambrian explosion hypothesis.

Future directions for accurately dating the eukaryote tree of life will be to increase the density of taxon sampling, to accommodate for among-protein heterogeneities in evolutionary rates, and to take into account the phylogenetic uncertainty [CR05].

References

[CR05] Cranston, K., and B. Rannala. 2005. Closing the gap between rocks and clocks. Heredity 94:461-462.

[DSB04] Douzery, E. J. P., E. A. Snell, E. Bapteste, F. Delsuc, and H. Philippe. 2004. The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils? *Proc. Natl. Acad. Sci. USA* 101:15386-15391.

[KTB01] Kishino, H., J. L. Thorne, and W. J. Bruno. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* 18:352-361.

[TKP98] Thorne, J. L., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* 15:1647-1657.

The Shapes of Phylogenies

Arne Ø. Mooers & Rutger Vos

Biological Sciences, Simon Fraser University, 8888 University Drive, V5A 1S6 - Burnaby BC - CANADA
amooers@sfu.ca

Species are quite inequitably distributed among evolutionary lineages [WY22]. This observation was first treated mathematically by Yule [Y24], who offered a very simple stochastic model that accounted for the known taxonomic patterns. One half of Yule's model (that of a single parameter governing the geometric growth of lineages) has become one of the two dominant null models for investigating the shapes of phylogenetic trees and so serves as a suitable departure point for a review of tree shape. There are at least four reasons to be interested in phylogenetic tree shape. First, the shape of the underlying tree might affect the ease with which we can correctly infer it. Second, the shape of the underlying tree might affect the ability to infer other parameters of interest (eg. ancestral states, correlated evolution). Third, the shape of the tree (or subtree) might inform us about the evolutionary (or ecological) forces that gave rise to it. Finally, the shape of a particular tree will affect how much of the tree is retained as species are pruned by extinction, a relationship of some interest in conservation biology and in predicting future diversification.

Phylogenetic tree shape is still considered on two separate axes [MH97][F04]. The first axis considers the topology, and ignores branchlengths. Though graphical representations of topology are possible [A01], it is common to calculate a single statistic that measures the variation in sizes of sister groups, usually with an eye to rejecting some model of diversification. Tree shapes are generally described as ranging from comblike (high variation in sister group sizes) to balanced (low variation in sister group sizes). The second axis measures the distribution of nodes from the root of the ultrametric tree to the tips, usually with an eye to fitting diversification parameters. Trees with their internal nodes nearer the tips are said to be stemmier than those with their internal nodes nearer the root. Stemmier trees can also be described as more redundant. For reasons that are not quite clear, these two ways of looking at a tree have rarely been considered together. One hypothesis worth investigating is that reasonable processes that give rise to comblike trees also give rise to less redundant trees.

We have known for at least 25 years that data evolved along certain tree shapes contain variation that makes the original tree difficult to reconstruct [F78]. Some recent simulation work suggests that such problematic shapes may not be uncommon (see, eg. [HL03]. In some cases, a more comblike tree may be harder to recover from the underlying gene trees that coalesce within it [R02] [DS05]. It is also possible that when a tree is reconstructed incorrectly, it is reconstructed to be more comblike than the true tree [F04]. This is hard to prove, however, since we do not generally know what shapes the underlying true trees actually have. It does seem to be the case that certain common consensus methods for producing "supertrees" may be biased towards producing comblike trees [WCC06].

Tree shape affects our ability to infer other parameters of interest primarily through its effect on redundancy. Trees with most of their nodes nearer the root sample more evolution, and so both should make ancestor reconstructions less precise and make measures of correlated evolution more precise. Subsampling from a larger focal tree might make trees less redundant than they otherwise would be.

Topologies (e.g. measures of balance) offer only coarse information about macroevolution. Differences between the topologies of the full tree and various subtrees (eg. of species sampled by ecological communities, by geographic regions, or by probability of extinction) may turn out to be more illuminating (see, eg. [WAM02]). In contrast, estimating diversification parameters from ultrametric trees is a powerful way to infer both tempo and mode of diversification, and the statistical framework [NMH94] is sophisticated (see [F04]). That said, producing ultrametric trees from real data is nontrivial, and it is conceivable that current methods are biased with respect to how nodes are distributed (e.g., see [HPC05].

Finally, while the Yule model of diversification is the most common starting point for investigating the properties of phylogenetic trees, the other standard model comes to us from the coalescent [H92]. Though the expected topologies produced by the two models is identical, the distribution of branch lengths is quite different, with the coalescent model producing trees with a great deal more redundancy. Yule models that incorporate extinction also increase redundancy; depending on the tempo of trait evolution, these models diminish any individual species' unique evolutionary contribution. From a conservation perspective, knowing how redundant real trees are, how that redundancy is distributed among tips, and whether current extinction is changing the redundancy of the subtrees that will remain are all of some importance.

References

[A01] D.J. Aldous, *Stat. Science* 16: 23-34, 2001.

[DS05] J.H. Degnan, L. A. Salter, *Evolution* 59: 24-37, 2005.

[F78] J. Felsenstein, *Syst. Zool.* 27: 401-410, 1978.

[F04] J. Felsenstein, *Inferring Phylogenies* (Sinauer Assoc., Sunderland, 2004)

[H92] J. Hey, *Evolution* 46: 627-640, 1992.

[HPC05] S.Y. W. Ho, M.J. Phillips, , A.J. Drummond, A. Cooper, *Mol. Biol. Evol.* 22:1355-1363, 2005.

[HL03] J.P. Huelsenbeck, K.M. Lander, *Syst. Biol.* 52:641-648, 2003.

[MH97] A.Ø. Mooers, S.B. Heard, *Q. Rev. Biol.* 72:31-54, 1997.

[NMH94] S. Nee, R.M. May, P.H. Harvey, *Phil. Trans. Roy. Soc. London B* 344:305-11, 1994.

[R03] N.A. Rosenberg, *Evolution* 57: 1465-1477, 2003.

[WCC06] M. Wilkinson, J.A. Cotton, C. Creevey et al., *Syst. Biol.* (in press)

[WAM02] C.O. Webb, D.D. Ackerly, M.A. McPeek, M.J. Donoghue, *Ann. Rev. Ecol. Syst.* 33:475-505, 2002.

[WY22] J. C. Willis, G.U. Yule. *Nature* 109:177-179, 1922.

[Y24] G.U. Yule, *Phil. Trans. Roy. Soc. London B* 213:21-87, 1924.

Bayesian Phylogenetics

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison
430 Lincoln Drive, Madison, WI, 53706-1381 — USA
brlarget@wisc.edu

Bayesian methods for phylogeny estimation are popular because they incorporate a likelihood-based approach to modeling genetic change through time and because assessment of uncertainty is computationally fast via Markov chain Monte Carlo (MCMC) relative to bootstrapping maximum likelihood. Another advantage of Bayesian methods is that interpretation of measures of uncertainty is straight-forward, albeit dependent on the specification of a prior distribution.

We will give an overview of the Bayesian approach to phylogenetics and then take a look at new directions in Bayesian phylogenetics research through several case studies.

Mitochondrial genome arrangements.— Most of the hundreds of animals whose mitochondrial genomes have been completely sequenced have the same set of 37 homologous genes, but the circular arrangements of genes differ. A Bayesian analysis of genome arrangements quantifies evidence for and against important evolutionary theories of early animal evolution [LSKS05].

AFLP evolution.— Amplified fragment length polymorphism (AFLP) markers are an increasingly common type of genetic marker. A Bayesian approach that models AFLP marker evolution as a hidden Markov process of the underlying DNA sequence evolution exploits fragment length information for improved inference [LHL05].

Concordance among multiple gene histories.— Rokas et al. [RWKC03] analyzed 106 genes from eight yeast species. An analysis of the combined data set results in complete confidence in a single tree, but there is substantial variation in optimal trees from single genes. We describe a Bayesian approach to combine separate single gene analyses with a prior distribution on gene tree clustering to assess the joint posterior distribution of gene trees.

Model averaging.— Analysis of aligned sequence data via Bayesian or maximum likelihood methods typically involves specification of a single likelihood model, such as JC69, HKY85, or the GTR model with gamma-distributed rates and invariant sites. It is common practice to test for a best model by some criterion and then carry out an analysis conditional on this model selection. The Bayesian approach offers the possibility of including the likelihood model selection as part of the model itself, effectively averaging over uncertainty in model selection [HLA04].

References

[HLA04] John P. Huelsenbeck, Bret Larget, and Michael Alfaro. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Molecular Biology and Evolution*, 21:1123–1133, 2004.

[LHL05] Ruiyan Luo, Andrew L. Hipp, and Bret Larget. A Bayesian model of AFLP marker evolution and phylogenetic inference. *Statistical Applications in Genetics and Molecular Biology*, 2005. in review.

[LSKS05] Bret Larget, Donald L. Simon, Joseph B. Kadane, and Deborah Sweet. A Bayesian analysis of metazoan mitochondrial genome arrangements. *Molecular Biology and Evolution*, 22:486–495, 2005.

[RWKC03] Antonis Rokas, Barry L. Williams, Nicole King, and Sean B. Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425:798–804, 2003.

Reconstructing Ancestral Recombination Graphs

Dan Gusfield

Department of Computer Science, Univeristy of California, Davis
gusfield@cs.ucdavis.edu
http://wwwcsif.cs.ucdavis.edu/ gusfield/

A phylogenetic network is a generalization of a phylogenetic tree, allowing structural properties that are not tree-like. With the growth of genomic data, much of which does not fit ideal tree models, and the increasing appreciation of the genomic role of such phenomena as recombination, recurrent and back mutation, horizontal gene transfer, gene conversion, and mobile genetic elements, there is greater need to understand the algorithmics and combinatorics of phylogenetic networks.

In this talk we survey a range of recent results on the algorithmics and combinatorics of phylogenetic networks with recombination. We discuss the problem of constructing a phylogenetic network for a given set of binary sequences derived from a known or unknown ancestral sequence, when each site in the sequence can change state at most once in the network, and recombination between sequences is allowed. The goal is to find a phylogenetic network that generates the given set of sequences, minimizing the number of recombination events used in the network. We show that when all the "recombination cycles" in the network are disjoint from each other, there are efficient (polynomial-time) algorithms that find a network minimizing the number of recombinations, and the optimal solution is "essentially unique". However, in general (when the cycles are not constrained), the problem is NP-hard, and in that case we present algorithms that are efficient in practice that obtain close upper and lower bounds on the number of recombinations needed. We also show a fundamental lower bound on the number of recombinations needed. Finally, we present a general decomposition theory about phylogenetic networks that shows the extent that the problem can be decomposed into a number of smaller problems, and we end with open question on that topic.

Joint work with Satish Eddhu, Chuck Langley, Dean Hickerson, Yun Song, Yufeng Wu and V. Bansal.

References

[GB05] D. Gusfield and V. Bansal. A fundamental decomposition theory for phylogenetic networks and incompatible characters. In *Proceedings of RECOMB 2005*, 2005.

[GEL04a] D. Gusfield, S. Eddhu, and C. Langley. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *Journal of Bioinformatics and Computational Biology*, 1:173–213, 2004.

[GEL04b] D. Gusfield, S. Eddhu, and C. Langley. The fine structure of galls in phylogenetic networks. *Informs J. on Computing special issue on Computational Biology*, 16:459–469, 2004.

[GHE05] D. Gusfield, D. Hickerson, and S. Eddhu. A fundamental, efficiently computed lower bound on the number of recombinations needed in a phylogenetic history. *Discrete Applied Math Special issue on Computational Biology*, 2005.

[Gus05] D. Gusfield. Optimal, efficient reconstruction of root-unknown phylogenetic networks with constrained recombination. *J. Computer and Systems Sciences, 2005 Special issue on Computational Biology*, 70:381–398, 2005.

[SWG05] Y. Song, Y. Wu, and D. Gusfield. Efficient computation of close lower and upper bounds on the minimum number of needed recombinations in the evolution of biological sequences. In *Proceedings of ISMB 2005*, 2005.

Groves – clustering phylogenetic databases.

Cécile Ané*, Oliver Eulenstein, Raul Piaggio-Talice and Michael J. Sanderson
*University of Wisconsin, ane@stat.wisc.edu

Supertree methods build a large tree by taking inputs consisting of smaller trees that share some taxa in common. Numerous real-world large-scale phylogenetic problems have now been addressed using supertree methods, including the phylogeny of all flowering plants, bacteria, mammals, and even dinosaurs. Yet there is little consensus about which of the many supertree methods perform best, or even satisfy specific properties, and a large number of basic theoretical questions remain unanswered.

We address here an issue first raised in a review of then existing supertree methods [SPH98] which has yet to be explored formally: what structure in the taxonomic overlap of the input trees is necessary for any supertree method to possibly reveal something new about phylogenetic relationships? Trivially, if two input trees share no taxa in common, then no supertree method will reveal something new about relationships that was not already known from the input trees taken separately. Sanderson et al. [SPH98] suggested that an overlap of two taxa between trees was necessary for the discovery of new information and proposed that "islands" of trees that overlap to this degree could serve as good substrates for supertree analysis. However, this idea was never formalized nor proven. In this work, we assess the potential informativeness of sets of input trees based on their taxon sets only, prior to actually applying any supertree method to them. The concepts we develop concern only the pattern of taxonomic overlap, not the agreement or disagreement of the input trees in regions of overlap.

We establish a criterion with which to cluster taxon sets. The clusters, called groves, have the potential to expose new information through supertree –or supermatrix– analysis. Loosely speaking, trees in a grove might contain new information about relationships from combined analysis, but the trees between two different groves do not (though they might contain new information with respect to the input trees within each of their respective groves).

We place upper and lower bounds on the minimum number of groves required to cover a set of input trees – or a set of input matrices. These bounds rely on overlap graphs for the taxon sets. Let $\mathcal{G}_1$ (resp. $\mathcal{G}_2$) be the overlap graph for the taxon sets where two taxon sets are connected by an edge if they share one (resp. two) taxa or more. We prove that the number of groves required to cover the database (the grove coverage number) lies between the number of connected components of $\mathcal{G}_1$ and the number of connected components of $\mathcal{G}_2$. We give criteria that allow us to determine the grove coverage number exactly in some cases. These results provide an answer to a question like how many supertrees must be built to include all the data in GenBank, for example (cf. [DAB$^+$04]).

References

[DAB$^+$04] A.C. Driskell, C. Ané, J.G. Burleigh, M.M. McMahon, B.C. O'Meara, and M.J. Sanderson. Prospects for building the tree of life from large sequence databases. *Science*, 306:1172–4, 2004.

[SPH98] M.J Sanderson, A. Purvis, and C. Henze. Phylogenetic supertrees: Assembling the trees of life. *Trends in Ecology and Evolution*, 13:105–109, 1998.

The mean, variance and limiting distribution of statistics sensitive to tree balance

Michaël Blum†‡ and Olivier François†

†TIMC-TIMB UMR CNRS 5525, Faculté de Médecine,
38706 La Tronche, France
‡Laboratoire Ecologie, Systématique et Evolution UMR CNRS 8079,
Bâtiment 360, Université Paris-Sud, 91405 Orsay, France
michael.blum@imag.fr, olivier.francois@imag.fr

The history of diversification rates among species leaves a strong footprint on the shape of the phylogeny that links extant taxa. A lot of attention has been paid to the way we can measure the shape of phylogenies. Simulation studies have investigated how these measures are sensitive to various macroevolutionary processes [KS93] . However, the probability distributions of these measures are not known theoretically except for a statistic called *the number of cherries* (i.e. the pairs of leaves that are adjacent to a common node) [MS00]. Using a connection with stream of works in theoretical computer science, we will give a probabilistic analysis for the Colless' and the Sackin's statistics which are the most common statistics used to capture tree balance [BF05].

Two probabilistic models of phylogeny will be considered: the Yule model which assumes that all species are equally likely to speciate, and the uniform model which assumes that all phylogenies are equally likely. In the Yule model, the computation of the limiting distributions is based on a contraction method which has been introduced for the probabilistic analysis of recursive algorithms [RR01] . In the uniform model, the computation of limiting distributions is based on the connection between uniform trees and uniform random walks. It turns out that the two statistics are strongly correlated in both models.

References

[BF05] M. Blum and O. François. On statistical tests of phylogenetic tree imbalance: The Sackin and other indices revisited. *Mathematical Biosciences*, 2005.

[KS93] M. Kirkpatrick and M. Slatkin. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution*, 47(4):1171–1181, 1993.

[MS00] A. McKenzie and M. Steel. Distributions of cherries for two models of trees. *Mathematical Biosciences*, 164:81–92, 2000.

[RR01] U. Rösler and L. Rüschendorf. The contraction method for recursive algorithms. *Algorithmica*, 29:3–33, 2001.

Syntheses of Life

Eric Bapteste, Dave Macleod, Robert L. Charlebois, Edward Susko and Ford Doolittle

GenomeAtlantic, 1721 Lower Water Street, Suite 401, Halifax, NS, B3J 1S5, CANADA
eric.bapteste@dal.ca,djmacleo@dal.ca, rlcharlebois@mac.com, susko@mathstat.dal.ca and ford@dal.ca

Lately, molecular phylogeny has seen a debate between the proponents of a tree of life and more skeptical evolutionists, such as Doolittle [D03][D00][D99a,b,c], Gogarten et al. [GDL02][ZLG04][ZG04], Lawrence[L99][L002], Lake[LR04] and Rivera [RL04] arguing that evolution was not tree-like. The debates rest on two main issues. First, the quality and quantity of phylogenetic signal is limited and this creates a challenging practical problem [BBLD04]. Second, a proportion of phylogenetic markers are incongruent with each other and support conflicting histories. To deal with this complex situation, we investigated the congruency of more than 400 phylogenetic markers from six datasets (gammaproteobacteria, alphaproteobacteria, bacteria, archaea, chloroplasts and eukaryotes) using heat map analyses. We conclude that most of the markers do not favor any unique tree, and even that they present some evidence for conflicting signal between them.

Based on these results, we tried to develop a critical phylogenetic approach which we suggest as a practical alternative to the flourishing supertrees methodologies. By reconstructing what we call Syntheses of Life [BBLD04], we try to depict molecular evolution with a concern of reconstructing only what can be reconstructed, without any *a priori* bias for or against lateral gene transfers (LGT). We do not aim to go for a classical Tree of Life, because genes may just be unable to provide us convincingly with such a broad coherent view. We feel that some parts of the molecular evolution are tree-like and some others are web-like, and that distinguishing these two structures matters. We argue that some understanding can be gained by respecting the contradictions in the data, and displaying lateral gene transfers events when describing molecular evolution in living beings, instead of hiding them under a majority rule [BW05].

Consequently, we have developed software [M05] that aims to extract evidence for vertical and lateral inheritance from a set of gene trees compared against an arbitrary reference tree. They provide important information to the evolutionary biologist such as the frequency and direction of putative LGT events and the robustness of a tree's backbone. This evidence is then displayed as a synthesis, a graph with both web-like and tree-like parts, showing support over the tree for vertical inheritance, overlaid with explicit lateral gene transfer events inferred to have occurred over the history of the tree. Finally, we applied our new phylogenetic drawing to the six datasets mentioned above. We also show how this representation could have heuristic value, allowing us to address questions such as the search of rules affecting lateral gene transfers events.

References

[D03] W. F. Doolittle et al. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? Philos Trans R Soc Lond B Biol Sci 358, 39-57; discussion 57-8, 2003.

[D00] W. F. Doolittle. Uprooting the tree of life. Sci Am 282, 90-5., 2000.

[D99a] W. F. Doolittle. Microbial evolution : The new synthesis. Proceedings of the 8th International Symposium on Microbial Ecology, 1-6, 1999.

[D99b] W. F. Doolittle. Lateral genomics. Trends Cell Biol 9, M5-8., 1999.

[D99c] W. F. Doolittle. Phylogenetic classification and the universal tree. Science 284, 2124-9, 1999.

[GDL02] J. P Gogarten, W. F. Doolittle and J. G. Lawrence. Prokaryotic evolution in light of gene

Reticulate Evolution

Charles Semple

Biomathematics Research Centre
Department of Mathematics and Statistics
University of Canterbury
Private Bag 4800
Christchurch, New Zealand
c.semple@math.canterbury.ac.nz

Evolutionary trees, also called phylogenetic trees, are used in evolutionary biology to represent the ancestral history of a collection of present-day species. However, because of reticulation events such as hybridizations and lateral gene transfers, evolution is not always tree-like. Consequently, instead of rooted trees, rooted acyclic digraphs are being used to model reticulate evolution, where vertices of out-degree zero represent the present-day species and vertices of in-degree at least two represent reticulation events. Such digraphs have a variety of names, but, for the purposes of this abstract, we will call them hybrid phylogenies.

Reticulation events are relatively rare, and so a fundamental problem for biologists studying the evolution of species whose past includes reticulation is the following: given an initial set of data that correctly represents the tree-like evolution of different parts of various species genomes, what is the smallest number of reticulation events required that simultaneously explains the entire data set. This smallest number sets a lower bound on the degree of reticulation that has occurred in the evolution of the species under consideration. Of course, associated with this problem is the algorithmic problem of reconstructing a hybrid phylogeny that explains the initial data set while simultaneously minimizing the number of reticulation events.

Typically, this problem has been studied from one of two viewpoints depending upon whether the initial data set is either a collection of phylogenetic trees or a collection of binary sequences (for example, see [BGMS04, GEL04, NWR04, WZZ01]). However, because the problem is NP-hard [BS05, WZZ01] (meaning that an efficient algorithm for solving it is unlikely to exist), much of the focus has been on identifying the mathematical structures that underlie this problem, developing fast algorithms for special cases of the problem, and providing ways to bound the smallest number of reticulation events.

In this talk, we will discuss some of the recent work that has been done on this problem particularly from the viewpoint when the initial data set is a collection of phylogenetic trees.

References

[BGMS04] M. Baroni, S. Grünewald, V. Moulton, and C. Semple, Bounding the number of hybridisation events for a consistent evolutionary history, *Journal of Mathematical Biology*, in press.

[BS05] M. Bordewich and C. Semple, Computing the minimum number of hybridisation events for a consistent evolutionary history, submitted.

[GEL04] D. Gusfield, S. Eddhu, and C. Langley, Optimal, efficient reconstruction of phylogenetic networks with constrained recombination, *Journal of Bioinformatics and Computational Biology* 2 (2004) 173-213.

[NWR04] L. Nakhleh, T. Warnow, and C. Randal Linder, Reconstructing reticulate evolution in species - theory and practice, in: *Proceedings of the 8th Annual International Conference on Research in Computational Molecular Biology* (RECOMB) (2004) 337-346.

[WZZ01] L. Wang, K. Zhang, and L. Zhang, Perfect phylogenetic networks with recombination, *Journal of Computational Biology* 8 (2001) 69-78.

Splits and Phylogenetic Networks

Daniel H. Huson

Center for Bioinformatics (ZBIT)
Tübingen University
72076 Tübingen, Germany
huson@informatik.uni-tuebingen.de

Phylogenetic trees are used to describe the evolution of a single gene under a model of evolution involving only mutation and speciation events. When studying the evolution of more than one gene, or under more realistic models of evolution, phylogenetic trees often do not suffice and more general phylogenetic networks must be employed. We give an introduction to a number of different types of phylogenetic networks, including*splits networks*, obtainable from sequences, distances or trees, *hybridization networks* obtainable from trees, and *recombination networks*, obtainable from binary sequences. One aim of this talk to show that these different types of networks are all closely related to each other. Further, we discuss a number of algorithms that compute such networks [DH04, HKLS05, HK05] and demonstrate their use on some examples. All presented concepts and algorithms are implemented in our program `SplitsTree4` [HB05].

References

[DH04] A. W. M. Dress and D. H. Huson. Constructing splits graphs. *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, 1(3):109–115, 2004.

[HB05] D. H. Huson and D. Bryant. Estimating phylogenetic trees and networks using SplitsTree 4. Manuscript in preparation, software available from `www.splitstree.org`, 2005.

[HK05] D.H. Huson and T.H. Kloepper. Computing recombination networks from binary sequences. To appear in: ECCB, 2005.

[HKLS05] D.H. Huson, T. Kloepper, P.J. Lockhart, and M.A. Steel. Reconstruction of reticulate networks from gene trees. In *Proceedings of the Ninth International Conference on Research in Computational Molecular Biology (RECOMB)*, 2005.

Poster Abstracts

Model Selection via Improved Phylogenetic Compression

Cécile Ané[1], Oliver Eulenstein[2], Raul Piaggio-Talice[2]

[1]University of Wisconsin, Madison, WI - USA
[2]Iowa State University, Ames, IA - USA
ane@stat.wisc.edu, oeulenst@cs.iastate.edu, rpiaggio@cs.iastate.edu

A central problem in phylogenetics is to select a hypothesis that best describes the evolutionary history leading to the sequences in a given alignment. Such a hypothesis can consist of a single tree for the whole alignment or multiple trees for different sections of it (as in the case of horizontal transfer or gene duplication). Ané and Sanderson recently proposed [AS05] a method of approaching this model selection problem by using the minimum description length principle from algorithmic information theory [LV93]. In this approach, the alignment is described by a two-part encoding composed of a code for a candidate hypothesis plus a code to recover the alignment given such hypothesis. The hypothesis assumed to be correct will be the one that minimizes the length of such encoding. Note that a minimum length encoding is also the best compression possible of the sequence alignment.

We present the mechanism proposed by [AS05] and propose a modification to it that results in provably shorter codes, closer to the minimum description length. The improvement is achieved by using ranking (and unranking, if the code is to be uncompressed) techniques. This is applied to code the hypothesis (tree or trees) as well as to code the sequence alignment given the hypothesis.

The shorter code provides a better compression mechanism and is expected to sharpen the hypothesis decision criterion. The new method still produces (efficiently) computable codes despite the fact that finding the hypothesis that minimizes a two-part code is akin to computing the Kolmogorov complexity of the alignment, known to be uncomputable [LV93].

When tested in real-world datasets from [SDEL03], the new method shows hypothesis distinction capabilities similar to that of the original version while the the compression improved on average by 26.18%, with gains that range from 8.79% to 49.02%.

References

[AS05] Cécile Ané and Michael J. Sanderson. Missing the forest for the trees: Phylogenetic compression and its implications for inferring complex evolutionary histories. *Systematic Biology*, 54(1):146–157, 2005.

[LV93] Ming Li and Paul Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, New York, revised and expanded second edition 1997 edition, 1993.

[SDEL03] Michael J. Sanderson, Amy C. Driskell, Oliver Eulenstein, and Sasha Langley. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. *Molecular Biology and Evolution*, 20:1036–1042, 2003.

Reconstructing Nucleotide Substitution Rates and Ancestral Nucleotide Distributions

Peter F Arndt

Max Planck Institute for Molecular Genetics
14195 Berlin, Germany
arndt@molgen.mpg.de

As already noted by Zuckerkandl and Pauling about 40 years ago, living natural systems preserve inscribed into their genomes the largest amount of their own evolutionary history [ZP65]. However, it is still a challenging task to accurately reconstruct this evolutionary history from present day sequence data. A very fruitful approach to this problem is the comparison of genomic sequences from different species and the reconstruction of their phylogeny using maximum likelihood methods [F04].

However, most of these approaches to phylogeny reconstruction make at least one of the following assumptions: (i) the substitution model is time-reversible and the same in all branches of a given tree (only the branch length might vary from one branch to another), (ii) not all substitution processes are considered independently, and (iii) the genomes under consideration are in the stationary state with respect to this model. These assumptions are sometimes necessary to efficiently compute the likelihood for a given substitution model and tree topology. However, these simplifying assumptions are not necessarily granted in reality. For vertebrates, they are actually violated to various degrees in the light of new sequence data and knowledge about the evolutionary processes [APH03, AH05].

Other work in this field does not make reference to the above assumptions. For example, Steel [S94] shows how one can recover a phylogenetic tree under the assumption that a simple Markov processes acted along the different branches. However, no method to reconstruct the underlying Markov process and the rates of the elementary substitution processes or the ancestral nucleotide distributions is known. Here we introduce such a method that is able to reconstruct all this ancestral information from present day sequence data only and which does not rely on the above mentioned assumptions. We will show how to reconstruct substitution frequencies under the most general 12 parameter model along the edges of a given phylogeny and ancestral nucleotide distributions using a maximum likelihood approach. Sequences at any node do not have to be in the stationary state and the underlying models may also be irreversible. While the position of the root cannot be fixed unambiguously, it can however be shown that substitution rates on branches not connected to the root and nucleotide distributions on all other nodes in the phylogeny are invariant with respect to the position of the root and can therefore be consistently reconstructed with the presented method.

References

[AH05] Arndt, P. F. and Hwa, T. (2005). Identification and Measurement of Neighbor Dependent Nucleotide Substitution Processes. *Bioinformatics* **21**: 2322-2328.

[APH03] Arndt, P. F., Petrov, D. A. and Hwa, T. (2003). Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. *Mol Biol Evol* **20**(11): 1887-96.

[F04] Felsenstein, J. (2004). Inferring phylogenies. Sunderland, Mass., Sinauer Associates.

[S94] Steel, M. (1994). Recovering a Tree from the Leaf Colourations It Generates under a Markov Model. *Applied Mathematics Letters* **7**(2): 19-23.

[ZP65] Zuckerkandl, E. and Pauling, L. (1965). Molecules as documents of evolutionary history. *J Theor Biol* **8**(2): 357-66.

Detection of heterotachy and its influence on phylogenetic inference within Eukaryotes

Guy Baele*, Yves Van de Peer+, Jeroen Raes+ and Stijn Vansteelandt*

*Department of Applied Mathematics and Computer Science
Krijgslaan 281 S9, 9000 - Gent - BELGIUM
+Department of Plant Systems Biology
Technologiepark 927, 9052 - Gent - BELGIUM
guy.baele@ugent.be

The substitution rate of a given site in a molecule is (1) not always constant through time and (2) can differ for different phylogenetic groups. Such behaviour is called heterotachy or within-site rate variation [LCP02]. Uncovering those positions that are heterotachous (i.e. which evolve according to the principle of heterotachy) is important for several reasons. First, it is a useful aid in the study of functional constraints. Second, it may deliver new insights that can help building more accurate evolutionary models for the reconstruction of phylogenetic trees. Finally, testing whether a given alignment contains heterotachous positions is important because the presence of such positions might disturb a reliable inference of phylogenetic relationships.

A principle for detecting heterotachy was proposed by Lopez, Casane and Philippe [LFP99] who verify for each site, using chi-square tests, whether the estimated number of substitutions across monophyletic groups is proportional to what can be expected based on the tree lengths. Because this procedure is sensitive to proportional changes of the tree lengths, we have developed a modified chi-square test. Further, to correct for multiple testing, we use recent ideas from Storey [ST03] by controlling the False Discovery Rate (FDR). This is defined as the proportion of false positives among all positions that are declared heterotachous by our testing procedure. Due to the small evolutionary rates, the procedure of Storey [ST03] was not directly applicable and a permutation-based modification was designed. Using permutations, we have also designed a method to determine which monophyletic groups in an alignment are primarily responsible for the heterotachy at a given position.

Our method detected heterotachy on 30% of all positions in a large eukaryotic dataset with sequences from 21 monophyletic groups. Mapping the heterotachous positions onto the secondary structure showed that the heterotachy is uniformly distributed along the structure as well as along the different regions of the structure (hairpin loops, stems, branching loops, internal loops and single stranded regions). This uniform pattern was also seen when mapping onto the tertiary structure. In the stem regions of the structure, we determined that in 71% the base pairs evolved according to a similar pattern (both according to heterotachy or not). This percentage is significantly elevated compared to the amount of co-evolution (or compensatory mutations) that would occur by chance.

Different test scenarios were constructed to determine the influence of the heterotachous positions on phylogenetic inference. Our attention primarily focused on the ability of the reconstruction method to correctly cluster the sequences in their respective monophyletic groups. Several groups showed a significant decrease in bootstrap support when the heterotachous positions were omitted, as compared to control experiments where an equal amount of random positions were omitted. Possible reasons for this trend will be discussed.

References

[LCP02] P. Lopez, D. Casane, and H. Philippe. Heterotachy, an important process in protein evolution. *Mol. Biol. Evol.*, 19(1):1–7, 2002.

[LFP99] P. Lopez, P. Forterre, and H. Philippe. The root of the tree of life in the light of the covarion model. *J. Mol. Evol.*, 49:496–508, 1999.

[ST03] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *PNAS*, 100(16):9440–9445, 2003.

An Application of Phylogeny in Genetic Epidemiology: the Detection of Disease Susceptibility Loci

Claire Bardel[1], Vincent Danjean, Emmanuelle Génin[1] and Pierre Darlu[1]

[1] INSERM U535 Génétique épidémiologique et structure des populations humaines
Hôpital Paul Brousse, Bâtiment Leriche, BP 1000, Villejuif, France
[2] LaBRI, équipe Runtime
UMR 5800 INRIA - Université de Bordeaux
351, cours de la Libration, 33405 Talence Cedex, France
bardel@vjf.inserm.fr, danjean@labri.fr, genin@vjf.inserm.fr, darlu@vjf.inserm.fr

The detection of an association between a gene and a disease and the localization of the susceptibility loci involved in the determinism of a disease are central questions in genetic epidemiology. Some classical methods involve the comparisons of the number of affected and non-affected individuals within the different categories of haplotypes existing in the population. However, the number of haplotypes may be large, thus decreasing the power of the test. In 1987, A.R. Templeton proposed a new method which consists in building a phylogeny of the haplotypes using a parsimony method and in performing the comparison of the number of affected and unaffected individuals within each clade [TBS87, Tem95]. If a clade contains an excess of affected individuals, one can conclude that mutations defining this clade could be involved in the disease.

This "cladistic method" has been applied on some real data sets by different authors, but its efficiency has never been assessed. In this work, we propose a new method to perform phylogeny-based association and localization analysis. In our method, a phylogenetic tree of the different haplotypes is reconstructed using parsimony or maximum likelihood methods and the ancestral character states are estimated. For each haplotype, a new character S is defined, the state of which depending on the proportion of cases and controls carrying this haplotype. Then, this new character S is adjusted on the phylogenetic tree and its correlated evolution with all the other sites is evaluated. We define the putative susceptibility sites as the sites which significantly co-mutates more often with S.

Using simulations, we evaluate the efficiency of this method for different genetic models. Our results show that when the susceptibility to the disease is due to only one locus, the phylogenetic analysis is not really more efficient than other classical methods. However, when two interacting sites are responsible for the disease, the phylogenetic analysis greatly improves the probability to find the two susceptibility sites. An application on a real data set concerning the CARD15 gene and Crohn disease shows that the method can successfully identify the three variant sites that are involved in the disease susceptibility. The effect of the phylogenetic method used to infer the haplotype evolution and to reconstruct the ancestral character states will also be discussed.

References

[TBS87] A R Templeton, E Boerwinkle, and C F Sing. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. *Genetics*, 117:343–351, 1987.

[Tem95] A R Templeton. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. analysis of case/control sampling designs: Alzheimer's disease and the Apolipoprotein E locus. *Genetics*, 140:403–409, 1995.

Hybrid phylogenies that reconcile two trees

Mihaela Baroni

Department of Mathematics
University Dunarea de Jos of Galati
str.Domneasca 47, 800008 - Galati - ROMANIA
mihaelabaroni@yahoo.com

To represent reticulate evolution, in particular hybridization, the model of hybrid phylogenies was proposed in [BSS04]. A fundamental problem is to determine the minimum number of hybrid events that are required to reconcile two incompatible gene trees with a consistent evolutionary history described by a hybrid phylogeny. As shown in [BGMS05], this minimum number h cannot be smaller than the rooted subtree prune and regraft (rSPR) distance between the two trees.

Recently, Bordewich and Semple [BS04, BS05] have proved that computing the rSPR distance and h are NP-hard problems. Confronted with the problem of computing h, one can try to reduce the problem to 'smaller' trees, by considering common clusters of the two trees. Some recent results will be presented.

References

[B04] M. Baroni, Hybrid phylogenies: a graph-based approach to represent reticulate evolution, PhD Thesis, University of Canterbury, Christchurch, New Zealand, 2004.

[BGMS05] M. Baroni, S. Grunewald, V. Moulton, and C. Semple, Bounding the number of hybridisation events for a consistent evolutionary history, *Journal of Mathematical Biology*, in press.

[BSS04] M. Baroni, C. Semple, and M. Steel, A framework for representing reticulate evolution, *Annals of Combinatorics* **8**(4):391-408, 2004.

[BSS05] M. Baroni, C. Semple, and M. Steel, Hybrids in real time, submitted

[BS04] M. Bordewich and C. Semple, On the computational complexity of the rooted subtree prune and regraft distance, *Annals of Combinatorics* **8**(4):409-423, 2004.

[BS05] M. Bordewich and C. Semple, Computing the minimum number of hybridisation events for a consistent evolutionary history, submitted

Phylogenetic invariants for the strand symmetric model

Marta Casanellas and Seth Sullivant

Departament Matemàtica Aplicada I. UPC
Av. Diagonal 647,
08028 Barcelona, SPAIN
marta.casanellas@upc.edu, seths@math.berkeley.edu

Given a phylogenetic tree and a model of base sequence evolution along the tree, a phylogenetic invariant is a polynomial that vanishes on the expected pattern frequencies at the terminal taxa. Knowing the phylogenetic invariants of a model for each possible tree topology allows one to infer the phylogenetic tree of a given set of taxa (see for example [PS05, Chapter 15]). Phylogenetic invariants were first introduced by Cavender and Felsenstein in [CF87], independently by Lake in [Lak87], and good references to the subject are [Fel03] and [PS05]. Recently many efforts have been made to find phylopenetic invariants: Allman and Rhodes got results for the general Markov model in [AR04] and Sturmfels and Sullivant [SS05] managed to describe all phylogenetic invariants for group-based models (Jukes-Cantor and Kimura 2 and 3 parameters).

In [PS05, Chapter 16] we have studied the phylogenetic invariants of strand symmetric Markov models. By a strand symetric Markov model on a tree we mean a Markov model whose mutation probabilities reflect the symmetry induced by the double-stranded structure of DNA. In particular in the root distribution we have $Prob(A) = Prob(T)$, $Prob(C) = Prob(G)$ and the substitution matrices associated to an edge i have the following form:

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} a_i & b_i & c_i & d_i \\ e_i & f_i & g_i & h_i \\ h_i & g_i & f_i & e_i \\ d_i & c_i & b_i & a_i \end{pmatrix} \end{array} .$$

Jukes-Cantor and Kimura models are special cases of strand symmetric models. A *general strand symmetric model* (SSM for short) is the one that has no more constraints on the substitution probabilites. The SSM captures more biologically meaningful features of real DNA data that the group-based models and it is not as general as the general Markov model, so we expect that knowing its phylogenetic invariants might lead to significant results on phylogenetic inference.

Our approach to describe phylogenetic invariants on the SSM for any tree is to extend the Fourier transform that had been used for group-based models. This extended Fourier transform has allowed us to compute the degree 3 and 4 phylogenetic invariants for the claw tree $K_{1,3}$ (see [PS05, section 16.2]) and to view the SSM on $K_{1,3}$ as a projection of the secant variety of a Segre variey.

Extending the arguments of [AR04] we deduce that the problem of determining defining phylogenetic invariants for the strand symmetric model on an arbitrary tree reduces to finding phylogenetic invariants for the claw tree $K_{1,3}$. Although we cannot give a complete list of invariants for $K_{1,3}$, we conjecture that all of them can be easily expressed as binomial relations on products of some matrices.

References

[AR04] E.S. Allman and J.A. Rhodes. Phylogenetic ideals and varieties for the general Markov model. Preprint, http://arxiv.org/abs/math.AG/0410604, 2004.

[CF87] J. Cavender and J. Felsenstein. Invariants of phylogenies in a simple case with discrete states. *Journal of Classification*, 4:57–71, 1987.

[Fel03] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Inc., 2003.

[Lak87] J.A. Lake. A rate-independent technique for analysis of nucleaic acid sequences: evolutionary parsimony. *Molecular Biology and Evolution*, 4:167–191, 1987.

[PS05] L. Pachter and B. Sturmfels, editors. *Algebraic Statistics for computational biology*. Cambride University Press, July 2005. ISBN 0-521-85700-7.

[SS05] B. Sturmfels and S. Sullivant. Toric ideals of phylogenetic invariants. *Journal of Computational Biology*, 12:204–228, 2005.

A General Method for Estimating the Transition Probability Matrix Under the GTR Model of DNA Sequence Evolution

Daniele Catanzaro[1], Raffaele Pesenti[2], and Michel C. Milinkovitch[1†]

dacatanz@ulb.ac.be - mcmilink@ulb.ac.be

Estimating nucleotide transition probabilities through time is at the core of many analytical methods in molecular evolutionary biology e.g., Distance Matrix Methods, Maximum Likelihood Approach, Invariants (see [FL2002]). Here, we formally characterize the biological and mathematical conditions that lead to the inapplicability of the general time reversible (GTR) model [LN1984] in general, and of the published estimation methods [WS1997, YG1996] in particular. We propose a new procedure that extends the set of conditions under which computing the transition probability matrix is feasible for the GTR model. This approach yields the transition probability matrix that maximizes the likelihood of observing the data. The analysis of the limitations of our extended GTR model opens perspectives regarding the development of alternative models.

References

[FL2002] Felsenstein, J. 2002. Inferring Phylogenies. Sinauer Associates, Sunderland, UK.

[LN1984] Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. Journal of Molecular Evolution 20:86-93.

[WS1997] Waddell, P. J. and M. A. Steel. 1997. General time reversible distances with unequal rates across sites: Mixing gamma and inverse gaussian distributions with invariant sites. Molecular Phylogenetics and Evolution 8:398-414.

[YG1996] Yang, Z. and S. Kumar. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rate among sites. Molecular Biology and Evolution 13(5):650-659.

[1] Laboratory of Evolutionary Genetics, Institute for Molecular Biology & Medicine (IBMM), Université Libre de Bruxelles, Rue Jeener et Brachet 12, B-6041, Gosselies, Belgium. www.ulb.ac.be/sciences/ueg

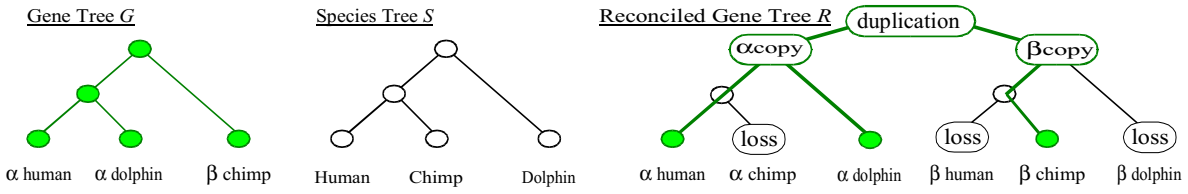
[2] Dipartimento di Ingegneria Informatica, University of Palermo, Viale delle Scienze I-90128 Palermo, Italy.

Gene-tree Reconciliation with Soft Multifurcations

Wen-Chieh Chang and Oliver Eulenstein

Iowa State University
Dept. of Computer Science
wcchang@iastate.edu, oeulenst@cs.iastate.edu

Species trees are implied from gene trees that represent evolutionary relationships of genes. Such species trees often differ from the true species tree, when implied from gene trees that are confounded by a complex history of gene duplication and losses.



The figure above gives an example; depicting a gene tree G derived from globins of human, dolphin and chimp and the true species tree S of these species. The gene tree G and species tree S differ, because of an ancient gene duplication taking place in the root species of S. Each copy of the duplication develops along the topology of the species tree S and results in the reconciled gene tree R. The gene tree G is a homeomorphic sub tree of R. Thus duplication is responsible for the incompatibility of trees G and S. The gene duplication could not be detected, because of the leaves in R that are not part of the embedding of G into R.

Gene duplication occurs in prokaryotes, but is especially common in eukaryotes, affecting as many as 40% of the genes, and leads to the diversification of entire gene families, i.e. globins or rhodopsins, in the same genome.

To reconcile a binary gene tree with a binary species tree Goodman et al. [GCM+79] introduced the model gene tree parsimony (GTP) in 1979. Later this model was refined and formalized by Page (1994), Guigo et al. (1996) and Eulenstein (1998). Reconciliations for the original and the extended GTP concepts can be computed in polynomial time (Zhang 1997, Eulenstein 1998).

A major shortcoming of the GTP extensions is that they were not designed to reconcile gene or species trees with 'soft multifurcations'. In practice gene and species trees are often inferred using 'soft multifurcations' to describe uncertainties. A *soft multifurcation* is a node in a rooted tree with more then two children that represents an unknown binary tree. None of the GTP extensions allows reconciling gene or species trees with soft multifurcations.

We extended GTP to reconcile gene and species trees with soft multifurcations, and developed a polynomial time reconciliation algorithm for our modified GTP model [Cha05]. The algorithm uses the dynamic programming paradigm to compute a reconciled tree, gene duplications and losses for the extended GTP concept. Briefly, our GTP model replaces every multifurcation by it's 'unknown tree' and then uses the GTP model to reconcile the modified gene tree. Following the parsimony approach the unknown trees are determined such that they minimize the size (number of nodes) of a reconciled tree. An implementation of the algorithm is in preparation and will be made publicly available upon completion.

References

[Cha05] Wen-Chieh Chang. Gene tree reconciliation with soft multifurcations. Master's thesis, Iowa State University, Dept. of Computer Science, 2005.

[GCM+79] M. Goodman, J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda. Fitting the gene lineage into its species lineage. a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, 28:132–163, 1979.

Choosing a method to detect and represent recombinations

Sylvain Goupil & Catherine Dauga

PF4, Génopole, Institut Pasteur, 28 rue du Dr Roux, 75724 Paris cedex 15
sgoupil@pasteur.fr, cdauga@pasteur.fr

Recombinations refer to several biological events : crossing-over, horizontal gene transfer and hybrid speciation. They occur at different levels : genome, population and species, that correspond to a large extent of mathematical situations, when DNA sequence comparisons are performed.

If the sequences are analyzed using phylogenetic tools that do not take recombination in account, misleading or incorrect trees are likely drawn, and mosaic structures might wrongly be ascribed to evolutionary forces other than recombination. Even low level recombination would have profound effects on phylogeny reconstructions [SH00]. So, the accurate detection of recombination from DNA sequences becomes very relevant and recently, new methods to represent relationships of mosaic sequences in phylogenetic networks have been implemented [H98].

Based on Phylogenetic- Substitution -Distance or Compatibility principles, methods available for detecting recombination can lead to conflicting results [P02]. However, only a few studies have attempted to examine their relative performance and, these studies were often limited in the set of conditions evaluated [PC01] [WCH01] [BGDJ01]. In the same way, when recombinations are present, the relationships between sequences may be illustrated by different types of phylogenetic networks (personal data).

We defined simulated data sets recovering a lot of evolutionary situations to understanding the performance of trees grafting into networks. By using a coalescent with recombination process [H83], we simulated recombinant genealogies. We focused on parameters frequently encountered in different biological situations : long sequences for genome exploration, large samples for population studies and low diversity for species characterization. We simulated also (reciprocal) recombinations and conversions (non reciprocal recombinations) of recent and ancient origins.

Recombination rates were estimated by using Plato and Recpars, two phylogenetic methods and Geneconv, a popular substitution method. The three tests detected equally (reciprocal) recombinations and old conversions but rarely conversions of recent origins. Their performance depends on variability level of sequences compared. False positives were observed especially with Recpars. False negatives were obtained with Geneconv and Plato when highly divergent sequences were compared, and with Recpars for closely related sequences.

The network tree algorithms were assessed by using SplitsTree, NeighborNet and ParsimonySplit on the different simulated data sets. All methods are suitable for long sequences (up to 10000bp tested). Only the method of decomposition of splits was usable for large samples (up to 500 sequences tested). The branch number of the NeighborNet trees increases exaggeratedly when the samples are getting larger.

Phylogenetic networks performed well with sequences of low variability and low recombination rates as well as high variability and high recombination rates. They represented networks for recombination and ancient conversions as well as recent conversions. Unfortunately, the three graph methods overestimated the presence of recombinations. False positives were observed even for samples without homoplasy. Among algorithms, ParsimonySplit gave the lowest number of false positives.

Network trees allow a considerable advance to represent non genealogical processes between genomes, populations or individuals within a species. They can be applied to a lot of biological situations. However, the graphs obtained in our study didnt reflect the amount of recombinations estimated between sequences.

References

[BGDJ01] CJ Brown, EC Garner, AK Dunker and P Joyce. The power to detect recombination using the coalescent. *Mol Biol Evol*, 18(7):1421-1424, 2001.

[H83] RR Hudson. Properties of the neutral allele model with intergenic recombination. *Theor Popul Biol*, 23:183-201, 1983.

[H98] DH Huson. Splitstree : analyzing and visualizing evolutionary data. *Bioinformatics*, 14:68-73, 1998.

[P02] D Posada. Evaluation of methods for detecting recombination from DNA sequences : empirical data. *Mol Biol Evol* : 19(5):708-717, 2002.

[PC01] D Posada and KA Crandall. Evaluation of methods for detecting recombination from DNA sequences : Computer simulations. *PNAS*, 98:13757- 13762, 2001.

[SH00] MH Schierup and J Hein. Consequences of recombination on traditional phylogenetic analysis. *Genetics*, 156:879-891, 2000.

[WCH01] C Wiuf, T Christensen and J Hein. A simulation study of the reliability of recombination detection methods. *Mol Biol Evol*, 18:1929-1939, 2001.

Graph Theory and Representation of Protein Dissimilarities

Gentian Gusho[1] [2]
[1]Département Logique des Usages, Sciences Sociales et de l'Information, ENST Bretagne,
Technôpole de Brest-Iroise (CS 83818), 29239 Brest cedex 3, France
[2]TAMCIC, U.M.R. CNRS 2872, ENST Bretagne
gentian.gusho@enst-bretagne.fr

There exist two main classification trends. The first is to approximate the data by a classificatory model: indexed hierarchies, pyramidal representations, weak hierarchies, while the second is to exctract clusters from the data as they are: the realisations (Brucker 2003 [B03]). In our case, clusters are extracted from a dissimilarity as *maximal cliques* of the *threshold graphs* of the dissimilarity. We associate a graph, with a minimum number of edges, to a *clustering system* such that every cluster be connexe on it. This graph is called *minimum rigidity graph* (MRG) and every restriction of such a graph, on the elements of any cluster, provides informations about the internal structure of the cluster and the main relations between its elements (pairs of elements connected by an edge).

The hierarchical model is a reliable model for biologists and the simple reason is a strong relation with the trees, supposedly representing the species evolution. For any protein dissimilarity, hierarchical representation is provided by the indexed hierarchy corresponding to the associated sub-dominant ultrametric. Ultrametrics which verify strong properties, are closely related with the *minimum spanning trees* (MSTs) of a dissimilarity (Leclerc 1981 [L81]). Indeed, every MST of a dissimilarity generates the associated sub-dominant ultrametric and rigidify its clusters. Moreover, they provide MRGs of the indexed hierarchy associated to the dissimilarity.

On the other hand, the sub-dominant ultrametric provides an approximation of the dissimilarity. The number of its clusters is linear while, generally, the number of clusters associated to a dissimilarity can be exponential. However, it is not necessary to study all the clusters but only a restricted number of them corresponding to the pairs of elements (Barthélemy 2003 [B03]). This brings us to a new definition of a dissimilarity cluster, the *realization* , which is defined from every pair of elements by the intersection of all the maximal cliques of the threshold graphs containing this pair. The new model of the *realizations* belonging to the second trend in classification, provides a broader representation of the dissimilarity information, compared to the hierarchical one. We show that an MRG of the realizations of a dissimilarity which always exists [B03], contains at least one of the MSTs associated to this dissimilarity.

We have shown these results in the case of a dissimilarity d representing the comparison scores between seven proteins of the family-16 of Glycosides Hydrolases. In the two representations of the Figure 1, only the MRG of the realizations shows the direct link between agarases 2 and 3 and κ-carrageenases 1 and 7 which belong to the same branch of the phylogenetic tree of the family.

d	2	3	4	5	6	7
1	88	88	86	85	87	73
2		69	88	88	89	85
3			85	80	80	88
4				10	15	82
5					16	87
6						86

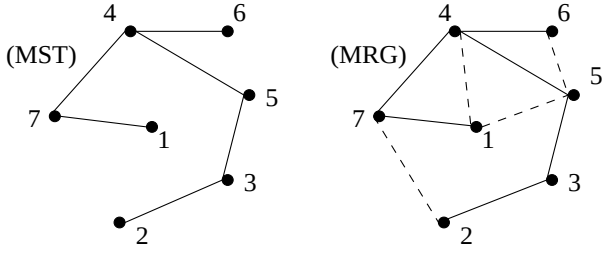


Figure 1: MST and MRG of the realizations of the dissimilarity d. In dotted lines are represented the edges that we add in the MST of d to obtain the MRG of the realizations of d.

References

[B03] F.Brucker, Réalisations de dissimilarités, *Actes des Rencontres de la Société Francophone de Classification*, 2003, pp. 7-10.

[L81] B.Leclerc, Déscription combinatoire des ultramétriques, *Mathématiques et Sciences Humaines*, **73** (1981), 5-37.

[B03] J.-P. Barthélemy, Classification binaire, *Actes des Rencontres de la Société Francophone de Classification*, 2003, 67-69.

Improving the Efficiency of SPR Moves in Phylogenetic Tree Search Methods Based on Maximum Likelihood

Wim Hordijk and Olivier Gascuel

Projet Méthodes et Algorithmes pour la Bioinformatique
LIRMM, UMR CNRS 5506 - Université Montpellier 2
161 rue Ada, 34392 - Montpellier - FRANCE
wim@santafe.edu, gascuel@lirmm.fr

Maximum likelihood (ML) methods are increasingly popular for constructing phylogenetic trees from sequence data. The main idea behind ML methods is that the space of possible trees is searched for an optimal topology (one that gives the maximum likelihood given the data), optimizing edge lengths along the way. Searching, or "moving", through this tree space is generally done using one of two methods: (1) local moves, such as nearest neighbor interchange (NNI), or (2) global moves, such as subtree pruning and regrafting (SPR). NNI moves can be implemented efficiently [GG03], but tend to get stuck on local optima. SPR moves, on the other hand, are less likely to get trapped on bad local optima and generally provide a more extensive search of the tree space [SLM04]. However, their main disadvantage is that they require a large amount of likelihood computation, making them much slower.

To improve current likelihood-based search methods, it is desirable to combine the advantages of both types of moves. We propose two methods to make SPR moves more efficient by reducing or even avoiding expensive likelihood computations. The first method uses a distance approach based on the minimum evolution principle which can be calculated very efficiently. Candidate SPR moves that are unpromising in terms of the minimum evolution criterion are simply discarded, which acts as a first filtering stage. This way, many unnecessary likelihood calculations can be avoided altogether. The second method involves updating a limited number of partial likelihoods (i.e., the likelihood values of subtrees) along the path between the prune and regraft positions, enabling a local calculation of the change in likelihood for candidate SPR moves as opposed to having to re-evaluate the entire tree. This reduces the amount of likelihood computation necessary, at only a small additional cost, for those remaining candidate SPR moves that were not already filtered out at the first stage.

We have implemented SPR moves together with the proposed efficiency-improving methods in the existing PHYML program [GG03], and applied it to data sets of the benchmark set used in [SLM04] for testing RAxML (another SPR based method). Our simulations show that, while indeed greatly reducing the amount of likelihood computation (by a factor of 40 to 60 compared to not using the efficiency improvements), the search results are on average as good as the best ones known so far. Table 1 shows the (average) likelihood values found by our SPR algorithm (alone (SPR) or in combination with NNI moves (PHYML+SPR)), compared to those of PHYML (without SPR moves) and RAxML.

	SPR		PHYML+SPR		PHYML		RAxML	
# taxa	random	parsim	random	parsim	random	parsim	random	parsim
101	-73869	-73874	-73878	-73869	-81851	-74004	-73926	-73870
150	-76856	-76856	-76855	-76858	-94099	-76925	-77161	-76853
250	-130920	-130940	-130902	-130920	-154791	-131083	-131325	-130897

Table 1: Likelihood scores found by our SPR algorithm compared to PHYML and RAxML. The results are averages over 10 random trees (random) or 10 parsimony trees (parsim).

References

[GG03] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704, 2003.

[SLM04] A. Stamatakis, T. Ludwig, and H. Meier. RAxML-III: A fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4):456–463, 2004.

A Taxonomy-traversing approach to discover cis-acting elements in Prokaryotes

Rekinś Janky and Jacques van Helden

SCMBB - Université Libre de Bruxelles.
Campus Plaine. CP 263. Boulevard du Triomphe. 1050 Bruxelles. Belgium.
[rekins, jvanheld]@scmbb.ulb.ac.be

The increasing number of sequenced genomes (more than 230 complete prokaryote genomes are now available on NCBI) opens promising avenues to apply comparative genomics in order to detect phylogenetically conserved cis-acting elements, and to study their divergence across taxonomy. This approach, called phylogenetic footprinting is based on the hypothesis that, due to selective pressure, regulatory elements tend to evolve at a slower rate than surrounding non-coding sequences. Many methods have being proposed, differing by their algorithm approaches and the way they take into account the phylogenetic relationship between the orthologs sequences. Phylogenetic footprinting has been applied to predict regulation in completely sequenced bacterial genomes [ALW04] [MTC+01] [MHC00]. Blanchette and Tompa [BST00] [BT02] [BT03] developed a dedicated algorithm, `Footprinter`, which take as input a set of upstream sequences and a taxonomic tree, and searches for conserved elements in each branch of the tree. The program gave impressive results for some examples from high organisms. However, motifs are scored according to parsimony criteria, and implicitly relies on an assumption of equiprobable and independent residues, which is not optimal for pattern discovery in microbial genomes. In order to decipher the evolution of transcriptional regulation in prokaryotes, we propose an original phylogenetic footprinting method to discover regulatory elements and to see their evolution in microbial genomes. Our strategy is to follow the prokaryotes taxonomic tree and to apply, at each branch, a pattern-discovery algorithm, called `dyad-analysis` [vHRCV00]. This algorithm performs a detection of over-represented dyads by comparison with a given background model. We illustrate the potentiality of our method taking as example lexA [EJS+04], a well-characterized bacterial gene involved in SOS response to DNA damage. This first analysis shows that motifs discovered by our method are consistent with several transcription factor binding sites annotated for different subtrees, or taxonomic classes (Gammaproteobacteria, Gram positive, Xanthomonadales). Interestingly, we also observe some divergence for a motif along distinct subclasses (Firmicutes and Actinobacteria). In the future, this approach will permit us to identify groups of co-regulated genes and to open the way for a better understanding of the evolution of transcriptional regulatory networks.

References

[ALW04] W. B. Alkema, B. Lenhard, and W. W. Wasserman. Regulog analysis: detection of conserved regulatory networks across bacteria: application to staphylococcus aureus. *Genome Res*, 14(7):1362–73, 2004.

[BST00] M. Blanchette, B. Schwikowski, and M. Tompa. An exact algorithm to identify motifs in orthologous sequences from multiple species. *Proc Int Conf Intell Syst Mol Biol*, 8:37–45, 2000.

[BT02] M. Blanchette and M. Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res*, 12(5):739–48, 2002.

[BT03] M. Blanchette and M. Tompa. Footprinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res*, 31(13):3840–2, 2003.

[EJS+04] I. Erill, M. Jara, N. Salvador, M. Escribano, S. Campoy, and J. Barbe. Differences in lexa regulon structure among proteobacteria through in vivo assisted comparative genomics. *Nucleic Acids Res*, 32(22):6617–26, 2004.

[MHC00] A. M. McGuire, J. D. Hughes, and G. M. Church. Conservation of dna regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res*, 10(6):744–57, 2000.

[MTC+01] L. McCue, W. Thompson, C. Carmack, M. P. Ryan, J. S. Liu, V. Derbyshire, and C. E. Lawrence. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res*, 29(3):774–82., 2001.

[vHRCV00] J. van Helden, A. F. Rios, and J. Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res*, 28(8):1808–18, 2000.

Continuous and Tractable Models for the Variation of Evolutionary Rates

*Thomas Lepage, †Stephan Lawi, *Paul Tupper and David *Bryant

*McGill Centre for Bioinformatics
3775 University, Montréal, Québec H3A 2B4 CANADA.
† Laboratoire de Probabilités et Modèles Aléatoires
Université Pierre et Marie Curie, 4, Place Jussieu, F-75252 Paris, FRANCE.

Understanding evolutionary rates and how they vary is one of the central concerns of molecular evolution. It has been clearly shown that inadequate models of rate variation, between lineages and between loci, can dramatically affect the accuracy of phylogenetic inference. The dependency of molecular dating on evolutionary rate models is even more critical: we will only obtain precise divergence time estimates from molecular data once we can model the rate at which sequences evolve.

Modelling the evolutionary rate is made difficult by the number and variety of factors influencing it. The base rate of mutation can vary because of changes in the accuracy of transcription machinery DNA repair mechanisms, and metabolic rate. At the cellular level, selective pressures can lead to variation of rate between loci and over time, as evidenced by differential rates of the three codon position, the slower evolutionary rate of highly expressed genes, and the effect of tertiary structure on patterns of sequence conservation.

Our goal is to derive a continuous model for rate evolution in order to avoid many of the problems of existing approaches (e.g. [KTB01]). We thouroughly examine the different aspects that need to be considered, mainly : 1) Continuity, 2) Long term behaviour and ergodicity, 3) Autocorrelation and the index of dispersion, and 4) Tractability.

Given these few natural criteria, we base our model on the CIR process, a continuous Markov process widely used in finance to model interest rates [CIR85]. This process appears as the simplest continuous model that is at the same time ergodic, has a non-zero autocovariance function and that can account for an arbitrarily large index of dispersion.

The model also fits well into existing protocols for phylogenetic inference. The process has a stationary distribution given by a gamma distribution and yet, unlike the rates-across-sites (RAS) model of Uzzell and Corbin, the rate is allowed to vary along lineages. The CIR model adds only one parameter to the RAS model, and this parameter can be estimated directly from the index of dispersion or the autocorrelation. Furthermore, this model is easily implementable in a MCMC framework. If we incorporate the model for evolutionary rate into the mutation model for sequence evolution at a site, these interact to give a joint process (R_t, X_t) for both the rate R_t at time t and the nucleotide or protein X_t at time t. To evaluate the likelihood we require an expression for the joint conditional probability

$$P[X_t = j, R_t = s | X_0 = i, R_0 = r]$$

of going from one nucleotide (or amino acid) state and rate state to another pair of states. In the CIR case, this probability has a closed form, and having an exact formula will speed up the computations significantly without having to resort to approximations, as in [KTB01]. The closed form is found by using the moment generating function of the integral (over time) of the process, in the same way Tuffley and Steel [TS98] derived a closed form for the transition probability for the Covarion model.

References

[CIR85] J. C. Cox, J. E. Ingersoll, and S. A. Ross. A theory of the term structure of interest rates. *Econometrica*, 53:385–408, 1985.

[KTB01] H. Kishino, J. L. Thorne, and W. J. Bruno. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution*, 18:352–361, 2001.

[TS98] C. Tuffley and M. A. Steel. Modeling the covarion hypothesis of nucleotide substitution. *Mathematical Biosciences*, 147:63–91, 1998.

Phylogenetic trees of minimal Shannon entropy

Christoph Mayer

Fakultät für Biologie
Lehrstuhl für Spezielle Zoologie
Gebäude ND 05/785
Ruhr Universität Bochum
44780 Bochum
Germany
cm@tp4.rub.de

One of the themes studied in information theory is the amount of information that is still left after a message has traversed a noisy channel. Clearly, the basic situation in information theory bears some similarity to the situation in molecular biology, where ancient DNA sequences, on their passage along internal edges of the true phylogenetic tree, can be envisaged as traversing noisy, but also more or less constrained, channels.

In molecular biology, our aim is to search for information in favor of possible phylogenetic trees, while having the problem that we do not know any of the ancestral sequences, and that we have a limited understanding of the mechanisms that operate on the sequences within the channels. Inspired by the fact that in information theory the information transmitted along a noisy channel is negatively related to the conditional entropy, we have defined a measure, based on the Shannon entropy, to determine the amount of information that a data set of aligned DNA sequences provides in favor of individual splits of a possible phylogenetic tree. The sum of these measures over all splits of a tree finally yield an approximate value of the amount of information in favor of the tree as a whole. This measure defines an optimality criterion for phylogenetic trees, which from its concept is related to the criterion of maximum parsimony.

We have tested our new optimality criterion on real and simulated data sets. The tests conducted so far indicate that the optimality criterion is well suited to recover the true phylogenetic trees on most reasonable data sets.

The following references provide a general introduction to Information Theory and entropy, where the first and last book focus on applications in biology. The above mentioned optimality criterion based on the Shannon entropy is still unpublished and is not described in these references.

References

[BW88] D.R Brooks and E.O. Wiley. Evolution as entropy, towards a unified theory of biology. The University of Chicago Press, Chicago, 1988.

[Kri86] K. Krippendorff. Information theory, structural models for qualitative data. Sage Publications, Newbury Park, 1986.

[Kul59] S. Kullback. Information theory and statistics. J. Wiley, New York, 1959.

[Yoc92] H. Yockey. Information theory and molecular biology. Cambridge University Press, Cambridge, 1992.

Site-Specific Evolutionary Rate Inference: Taking Phylogenetic Uncertainty into Account

Itay Mayrose and Tal Pupko

Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences
Tel Aviv University, Ramat Aviv 69978, Israel
{itaymay,talp}@post.tau.ac.il

The degree to which an amino acid site is free to vary is strongly dependent on its structural and functional importance. An amino acid that plays an essential role, such as one within the active site of the protein, is unlikely to change over evolutionary time. Hence, the evolutionary rate at an amino-acid site is indicative of how conserved this site is, and in turn, allows evaluating the importance of this site in maintaining the structure or function of a protein.

Accurate inference of evolutionary rates requires the reconstruction of a phylogenetic tree describing the evolutionary relationship among the sequences under study. The general algorithm for rate inference described to-date is composed of two basic steps: (1) constructing the phylogenetic tree, and (2) inferring site-specific rates based on this phylogenetic tree. Various inference methods differ in the manner step (2) is performed, while all rely on the assumption that the phylogeny is absolutely correct. Such a naive algorithm has the potential to result in erroneous estimates, if the inferred phylogeny does not reflect the reality. When the combined state $\omega = \{\tau, t, \theta\}$ is unknown we would ideally consider all possible tree topologies τ, branch lengths t, and model parameters θ. However, this approach is intractable for realistic sized problems. We developed a novel Bayesian method, McRate [MMP05], that uses Markov-chain-Monte-Carlo methodology [MRR$^+$53] to generate a large sample from the posterior probability distribution of states without explicit enumerating this complicated space. This method computes site-specific rates taking into account alternative tree topologies, branch lengths, and model parameters.

Simulations were used in order to compare the accuracy of the rates inferred by McRate and those inferred by an empirical Bayesian method [MGBTP04], in which inference is based on a single tree (ST). Our simulations tested different model trees and different rate distributions. ST requires for its computations a given phylogenetic tree. Two different tree reconstruction algorithms were examined: maximum likelihood (ML), using the SEMPHY program [FNPP02] and neighbor-joining [SN87], referred to as ST-ML and ST-NJ, respectively. A comparison between the inference accuracy of McRate and the two ST methods for different number of sequences is shown in Table 1. Our results indicate that McRate is the most accurate method while ST-NJ seems the least accurate one.

# of sequences	MSE[a] McRate	MSE ST-ML	MSE ST-NJ	P value[b] McRate vs. ST-ML (ST-NJ)
7	0.1046	0.1064	0.1091	0.0196 (0.0003)
17	0.0836	0.0859	0.0912	0.12 (0.0007)
27	0.0717	0.0762	0.0816	0.037 (0.0008)

Table 1: Simulation results: Model trees with different number of sequences. Simulated rates were drawn from a Γ distribution with $\alpha = 0.3$. [a]MSE is the average mean square error obtained over 30 independent runs. [b]P value was calculated using Wilcoxon nonparametric test between two dependent samples.

References

[FNPP02] N. Friedman, M. Ninio, I. Pe'er, and T. Pupko. A structural em algorithm for phylogenetic inference. *J Comput Biol*, 9(2):331–353, 2002.

[MGBTP04] I. Mayrose, D. Graur, N. Ben-Tal, and T. Pupko. Comparison of site-specific rate-inference methods for protein sequences: empirical bayesian methods are superior. *Mol Biol Evol*, 21(9):1781–1791, 2004.

[MMP05] I. Mayrose, A. Mitchell, and T. Pupko. Site-specific evolutionary rate inference: taking phylogenetic uncertainty into account. *J Mol Evol*, 60(3):345–353, 2005.

[MRR$^+$53] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J Chem Phys*, 21:1087–1092, 1953.

[SN87] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, 1987.

A multi-gene approach to resolve molecular phylogeny of Malagasy dung beetles

L. Orsini, H. Koivulehto and I. Hanski

Metapopulation Research Group
Department of Biological and Environmental Sciences
PO Box 65 (Viikinkaari 9) – FI-00014 University of Helsinki – FINLAND
helena.koivulehto@helsinki.fi

Madagascar has been isolated from Africa for 160 and from India for 90 million years [DW03] and is unique for the high degree of endemism at high taxonomic levels and for skewed representation of many groups of plants and animals. Dung beetles (Coleoptera, Scarabaeinae) have radiated in Madagascar in close association with lemurs (Lemuriformes, Primates), the largest herbivores on the island.
The endemic Malagasy dung beetles belong to the tribes Helictopleurini (Coprinae) and Canthonini (Scarabaeinae) [L60,C91], with one genus and ~80 species and 12 genera and ~170 species, respectively. We have obtained a large fraction of the species in an extensive trapping program across Madagascar.

One of the most pervasive challenges in molecular phylogenetics is the incongruence of phylogenies obtained using different data sets based on different individual genes. To resolve the phylogeny of Malagasy dung beetles we screened the sequences of 2 nuclear (18S and 28S) and 4 mitochondrial genes (12S, 16S, COI, Cytb) singly and by concatenation. Similar results were obtained for single genes and while concatenating them. The two classes of genes show similar patterns, with a higher resolution for species in the phylogeny based on mtDNA sequences.
Helictopleurini represents a monophyletic group, within which two major groups are evident. The *giganteus* group of large species is well supported by distance, parsimony, and likelihood analyses and suggests that the classification based on morphological characters is correct. Within this monophyletic group there are no basal species present.
Five genera of Canthonini have been included in the analysis so far. The genus *Aleiantus* is in a basal position, though this result is not statistically well supported. *Aleiantus* and *Sphaerocanthon* are monophyletic while *Apotolamprus* and *Arachnodes* appear to be polyphyletic, with species in two different clades. This result may have two explanations: 1) the taxonomic classification based on morphology is not correct; 2) parallel radiation has taken place among species belonging to the two genera.
Simultaneous peripatric and sympatric speciation events are likely to produce polytomies and lack of phylogenetic resolution is suggestive of simultaneous peripatric or parapatric speciation events. Our phylogenies lack statistical support when the nuclear markers are used and polytomies are present in trees built with nuclear and mitochondrial markers. This result agrees with a hypothesis of rapid speciation in this group of species. This phylogenetic analysis will be placed into an evolutionary context in which traces of selection acting on the genome will be investigated.

References

[DW03] De Wit, J.M. 2003: Madagascar: Heads it´s a continent, tails it´s an island — Annual Review of Earth Planet Sciences 31:213–248

[L60] Lebis, . 1960: Faune de Madagascar Insectes Coléoptères Scarabaeidae, Helictopleurina — Publications de l´Institut de recherche scientifique, Tananarive – Tsimbazaza

[C91] Cambefort, Y. 1991: Biogeography and evolution — pp 51–68 in Dung Beetle Ecology, I. Hanski & Y. Cambefort eds. Princeton University Press

Tracing the origin of the HIV-1 CRF04_cpx initially designated as "subtype I" in Greece using Bayesian method

D. Paraskevis, E. Magiorkinis, G. Magiorkinis, A. Hatzakis
National Retrovirus Reference Center, Department of Hygiene and Epidemiology, Athens University Medical School, Greece
dparask@cc.uoa.gr

CRF04_cpx, initially designated subtype I, is one of the 16 HIV-1 circulating recombinant forms representing recombinant HIV-1 genomes. CRF04_cpx was documented in Cyprus and Greece and it was found to be comprised of at least five distinct subtypes (A, G, H, K and unclassified regions) [GRCL98+] [NPMT99+] [PMVK01+].

To estimate the origin of the CRF04_cpx epidemic using all available CRF04_cpx sequences. The most recent common ancestor (MRCA) of the CRF04_cpx cluster was estimated for 10 sequences, sampled from 9 individuals, in partial *gag* and partial *env* regions. The estimation of the substitution rate was performed by inferring simultaneously population, substitution parameters and tree topology by using Bayesian inference as implemented in BEAST (v1.0.3) [DR03] [DNRS02] and using a different substitution and evolutionary model (GTR+Γ) for *gag* and *env* partitions. Three separate MCMC runs were made for 3 X 10^6 generations with a burnin of 3 X 10^5.

The mean substitution rate (combined runs) for *gag* and *env* was estimated 2.1 X 10^{-3} [95% highest posterior density interval, (HPD): 1.25 X 10^{-3} - 2.98 X 10^{-3}], and 5.66 X 10^{-3} [95% HPD: 3.42 X 10^{-3} - 7.92 X10^{-3}], substitutions per site per year, respectively. The mean dates of the most recent common ancestor (MRCA) of the tree (t_{root}) was estimated in 1973.9 [95% HPD: 1964.4 – 1981.5]. Interestingly, for 5 patients infected by a common source between 1989 and 1993 in Thessalonica Northern Greece, the coalescent event was in 1985.6 [95% HPD: 1980.7 – 1989.4] [PPPK02+] [PMMK04+].

The MRCA of the CRF04_cpx sequences in 1973.9 precedes the date of the earliest known HIV-1 infections in the early 1980s in Greece, suggesting probably that CRF04_cpx originated outside Greece. Interestingly, the origin of the CRF04_cpx dates only a few years later than the date of origin of the US HIV-1 subtype B epidemic, thus suggesting that it originated at the early years of the global HIV-1 epidemic spread.

References

[DR03] Drummond AJ, and Rambaut A. 2003 BEAST v1.0, available from http://evolve.zoo.ox.ac.uk/beast/.

[DNRS02] Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 2002; 161:1307-1320.

[GRCL98+] Gao F, Robertson DL, Carruthers CD, Li Y, Bailes E, Kostrikis LG, Salminen MO, Bibollet-Ruche F, Peeters M, Ho DD, Shaw GM, Sharp PM, and Hahn BH. An isolate of human immunodeficiency virus type I represents a complex mosaic comprising three different group M subtypes (A, G, and I). J Virol 1998; 72:10234-10241.

[NPMT99+] Nasioulas G, Paraskevis D, Magiorkinis E, Theodoridou M, and Hatzakis A. Molecular analysis of the full-length genome of the HIV-1 "Subtype I": evidence of triple recombination. AIDS Res Hum Retroviruses 1999; 15:745-758.

[PPPK02+] Pappa, A., E. Papadimitriou, A. Papoutsi, V. Kiosses, and A. Antoniadis. 2002. HIV-1 subtypes and circulating recombinant forms (CRFs) in nortern Greece. Virus Res. 85:85-93.

[PMMK04+] Paraskevis D, Magiorkinis E, Magiorkinis G,. Kiosses V.G., Lemey P, Vandamme A-M, Rambaut A, Hatzakis A. Phylogenetic reconstruction of a known HIV-1 CRF04_cpx transmission network using maximum likelihood and Bayesian methods. J Mol Evol 2004; 59:709-717.

[PMVK01+] Paraskevis D, Magiorkinis M, Vandamme A-M, Kostrikis LG, and Hatzakis A (2001) Re-analysis of human immunodeficiency virus type 1 isolates from Cyprus and Greece, initially designated as 'subtype I', reveals a unique complex A/G/H/K/? mosaic pattern. Journal of General Virology 82:575-580.

Inferring Phylogenies by Confidence Set Optimization

S. L. Pepke, Davin Butt, Isabelle Nadeau, Andrew J. Roger, and Christian Blouin

Genome Atlantic/Dalhousie University
Halifax, NS, CANADA
pepke@hades.biochem.dal.ca, Andrew.Roger@dal.ca, cblouin@cs.dal.ca

The problem of searching effectively in tree space for the maximum likelihood tree becomes a formidable one as the number of possible trees grows rapidly with the number of taxa. Search is typically performed via systematic rearrangement of the current best tree estimate. For a greedy algorithm, search is halted when none of the topological rearrangements yield a tree with greater likelihood than the current best estimate. Clearly, the maximum discovered in this way may be only local. We examine the impact of local likelihood maxima on the accuracy of phylogenetic inference. For both real and simulated data we find that local maxima are present and hinder convergence to the maximum likelihood tree when single-threaded search is used. However, when pools of trees that exhibit likelihoods within a narrow window of the current maximum are utilized during the search, the rate of finding the global maximum likelihood tree is greatly improved.

We simulate data on a variety of topologies to investigate the relationship of the estimated maximum likelihood tree to the true tree and find that when topologies contain very short internal branches and/or long terminal branches, the maximum likelihood tree is often not the true tree. In contrast, when confidence sets are used to determine the candidate pool during search, the probability of the final confidence set containing the true tree can be very high. We compare various confidence set definitions, including those of Shimodaira and Hasegawa [SH99] and Strimmer and Rambaut, as well as others. [SR02]. The results shown in figure 1 indicate that for challenging inference problems, it is advisable to use the relatively conservative SH test to ensure sufficient breadth of search for convergence to a set containing the true topology. We also see that even in cases where the confidence set reliably contains the true tree, the confidence set consensus tree is not likely to be equal to the true tree. Thus one should be cautious in interpreting consensus trees. When breadth search based upon confidence sets is applied to real protein sequence alignments including mammal mitochondrial proteins and an alpha tubulin alignment of microsporidia and ascomycetes, results appear consistent with that for the simulated sequences.

In summary, phylogenetic inference can be made more reliable by (1) improving search in topology space through breadth search utilizing a pool of candidate ML trees, and (2) choosing inference methods (confidence sets) that are well matched to the information content of the sequence data.

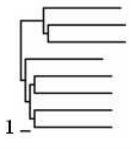
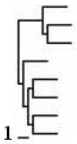
Topology	$L_{internal}$	$L_{external}$	Test	True Found (%)	True is ML (%)	Consensus is True (%)	Average Confidence Set Size
	.05	.05	RELL	100	47	37	316
			SH	99	44	30	57
			ELW	62	75	31	2.8
	.05	.25	RELL	100	75	18	286
			SH	100	82	64	14
			ELW	93	88	69	1.6

Figure 1: Relationship of maximum likelihood, confidence set consensus, and true topologies for simulated sequence data.

References

[SH99] Hidetoshi Shimodaira and Masami Hasegawa. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.*, 16(8):1114–1116, 1999.

[SR02] Korbinian Strimmer and Andrew Rambaut. Inferring confidence sets of possibly misspecified gene trees. *Proc. R. Soc. Lond. B*, 269:137–142, 2002.

Multivariate Analysis for Data Collected on Species

Elizabeth Purdom and Susan Holmes

Department of Statistics
Stanford University
390 Serra Mall, Stanford, CA 94305-4065
epurdom@stanford.edu,susan@stat.stanford.edu

It is a common occurrence in biology to collect data on many different species, such as morphological traits, protein/DNA sequences, or geographic proximity. For instance, in traditional ecological studies the data collected for each species is the abundance or presence/absence of the species in the different geographic sites of interest. A large array of statistical techniques have been developed to analyze such ecological data, e.g. correspondence analysis. The general statistical question, in some form, is to evaluate the relationship between the different variables, such as the sites, that are observed for each species. Fundamentally this requires appropriate methods of measuring the similarity of different traits.

Unlike the usual scenario in statistics, however, the observations, i.e. species, cannot be assumed to be independent observations from the same distribution. The phylogenetic relationship between the species creates possible dependence among the observations. Usual measures of similarity will be confounded with the residual phylogenetic relationships of the species. This will be particularly noticeable when there is an unequal representation of different species across the tree.

A recent method called Double Principal Coordinate Analysis (DPCoA) [PDC04] compares different sites using Rao Diversity/Dissimilarity measurements, which incorporates measures of phylogenetic distance between the species. We demonstrate the use of DPCoA on human intestinal microbial flora and its effectiveness in analyzing microbial ecological studies. The Rao Dissimilarity attempts to measure the underlying dissimilarity/diversity of populations and is not explicitly attempting to correct for the statistical dependence of the species. Indeed, the distances between species do not have to come from a phylogenetic tree.

When the dissimilarities between species are based on a phylogenetic tree, the DPCoA of ecological data complements many comparative methods for analyzing traits observed on species. In particular, DPCoA can be seen as a principal components analysis of a space with a different inner product space than the traditional Euclidean space. This same space is connected to traditional comparative methods used in phylogenetic studies of continuous data, such as phylogenetically independent contrasts [Fel85], or more generally the generalized least squares method [Gra89]. These comparative methods use a covariance matrix resulting from traits following a Brownian motion process along the tree; this covariance matrix is linearly related to the patristic distance between traits.

Use of different inner products for traits can be generalized to abstract spaces through kernel methods. Such methods rely on evaluating inner products of observed objects on very general spaces. We demonstrate how the inner product structure in comparative methods can be related to some kernel methods for comparing data on species.

The analysis of the microbial flora presented can be found in [EBB+05]. This work has been supported by a National Science Foundation grant, DMS 02-41246.

References

[EBB+05] Paul B. Eckburg, Elisabeth M. Bik, Charles N. Bernstein, Elizabeth Purdom, Les Dethlefsen, Michael Sargent, Steven R. Gill, Karen E. Nelson, and David A. Relman. Diversity of the human intestinal microbial flora. *Science Express*, 14 April 2005.

[Fel85] Joseph Felsenstein. Phylogenies and the comparative method. *American Naturalist*, 125:1–15, 1985.

[Gra89] A. Grafen. The phylogenetic regression. *Phil. Trans. Royal Society of London, Series B*, 326:119–157, 1989.

[PDC04] Sandrine Pavoine, Anne-Béatrice Dufour, and David Chessel. From dissimilarities among species to dissimilarities among sites: a double principal coordinate analysis. *Journal of Theoretical Biology*, 228:523–537, 2004.

Site-interdependent models of sequence evolution: using statistical potentials in phylogenetic analyses

Nicolas Rodrigue [1], David Bryant [2], Hervé Philippe [1], Nicolas Lartillot [3]

1. Canadian Institute for Advanced Research
Département de biochimie, Université de Montréal
2900 Édouard-Montpetit - Montréal, Québec - CANADA
nicolas.rodrigue@umontreal.ca
herve.philippe@umontreal.ca

2. McGill Centre for Bioinformatics
Montreal, Québec - CANADA
bryant@mcb.mcgill.ca

3. LIRMM
Montpellier - FRANCE
nicolas.lartillot@lirmm.fr

Most models of sequence evolution applied in phylogenetics today assume independence between sites. This computationally motivated simplification is known to be biologically unsound. We have recently applied the sampling methods described in [RJK+03] to propose a model that explicitly accounts for site-interdependencies resulting from protein tertiary structure [RLBP05]. The model is based on statistical potentials, which have typically been studied in the context of protein fold prediction and protein structure model validation. We use these potentials as sequence fitness proxies, to study the importance of protein structure in shaping the evolutionary process. We have implemented our model in a Bayesian MCMC framework. Through numerical evaluations of Bayes factors, we show that site-interdependence due to protein tertiary structure is always favored over standard models for all datasets studied. We also show how our techniques can be used to compare different statistical potentials in this context, and how to best combine them with standard site-independent models. Although focusing on protein tertiary structure, the techniques we present should allow for the study of a broader range of models, including any site-interdependent criteria.

References

[RJK+03] D. M. Robinson, D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. Protein evolution with dependence among codons due to tertiary structure. *Mol. Biol. Evol.*, 18:1692–1704, 2003.

[RLBP05] N. Rodrigue, N. Lartillot, D. Bryant, and H. Philippe. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene*, 347:207–217, 2005.

Sequence diversity at multiple, conserved housekeeping genes by analyzing Single Nucleotide Polymorphism for determining the population structure and molecular evolution of *Salmonella* Typhi

Philippe Roumagnac[1], Sylvain Brisse[2], Gordon Dougan[3] and Mark Achtman[1]

[1]Department of Molecular Biology, Max-Planck-Institut fr Infektionsbiologie, Schumannstrasse 21/22, 10117 Berlin, Germany, roumagnac@mpiib-berlin.mpg.de

[2]Unit Biodiversit des Bactries Pathognes Emergentes, Institut Pasteur, 25-28 rue du Docteur Roux, 75724 Paris cedex 15

[3]Center for Molecular Microbiology and Infection, Department of Biological Sciences, Imperial College London, London SW7 2AZ, UK

Salmonella enterica subspecies *enterica* serovar Typhi (*S.* Typhi) causes typhoid fever in humans. Approximately 16-20 million cases occur annually, resulting in 600,000 deaths [PLI+98]. *S.* Typhi belongs to a group of bacterial species that are so young that they have not yet had sufficient time to accumulate sequence diversity [KRW+02], including *Yersinia pestis*, *Mycobacterium tuberculosis* and *Plasmodium falciparum*. Hence, the analysis of population structure in such organisms is difficult because sequence polymorphisms are so rare. We have therefore performed extensive screening of multiple genes from a globally representative collection of *S.* Typhi in order to accumulate sufficient data. To this end, we have used the dHPLC method to identify single nucleotide polymorphisms (SNPs) within 81 selectively neutral housekeeping genes among a world-wide collection of *S.* Typhi strains. Such selectively neutral SNPs can be used for assessing the number of haplotypes as well as their frequency distribution within the population in order to measure genetic diversity. Currently, we have identified 36 SNPs (24 are synonymous) which define 27 haplotypes. Nineteen haplotypes are represented by only a single strain, six contain two to seven strains, one contains 13 and another contains 45. These two latter common haplotypes differ from each other by only one synonymous SNP and form the centers of two radial expansions of haplotypes differing from their parent by only one to five SNPs. The results subdivide *S.* Typhi into two widespread clusters of haplotypes, which can now be matched against patterns of epidemic spread. The data will be used to estimate important parameters of the population structure of *S.* Typhi including the ratio of recombination to mutation, importance of selection, and the effective population size.

References

[KRW+02] C. Kidgell, U. Reichard, J. Wain, B. Linz, M. Torpdahl, G. Dougan, and M. Achtman. *Salmonella typhi*, the causative agent of typhoid fever, is approximately 50,000 years old. *Infect. Genet. Evol.*, 2:39–45, 2002.

[PLI+98] T. Pang, M. M. Levine, B. Ivanoff, J. Wain, and B. B. Finlay. Typhoid fever—important issues still remain. *Trends Microbiol.*, 6:131–133, 1998.

A Novel Way for Combined Phylogenetic Analysis between Supertrees and Supermatrices

Heiko A. Schmidt[1] and Arndt von Haeseler[1,2]

[1] Bioinformatics, NIC, FZ Jülich, Germany
[2] Bioinformatics, HHU Düsseldorf, Germany
hschmidt@cs.uni-duesseldorf.de, haeseler@cs.uni-duesseldorf.de

Although the amount of sequence data in the primary sequence databases like EMBL and GenBank doubles within about nine months, mutual coverage of taxa and genes is far from being satisfactory. This lack of coverage poses serious problems when reconstructing trees for larger datasets due to substantial amounts of missing data.

Mainly two ways have been suggested to overcome this problem. First, supermatrix or 'total evidence' approaches [Klu89] combine the available alignments into one large supermatrix, from which one tree is reconstructed (early combination).

The second approach constructs so-called supertrees [BE04 for overview] from a set of trees reconstructed separately from the available datasets (late combination) using agreement or common nestings among the input trees or by decomposing the trees into, e.g., quartets [RRG01] or matrix representation BE04 to construct the supertree.

Here we suggest a third (medium) level of combining the available data for a phylogenetic reconstruction. To this end we introduce the overlap graph to measure the 'combinability' of the available datasets. This graph points to crucial lacks of overlapping information among the datasets and serves as a guide during the reconstruction process.

For the actual reconstruction method we evaluate all possible quartet topologies for each dataset, e.g., using maximum likelihood [ML, Fel81]. For a given quartet of taxa the likelihood of the quartet trees from the available sequence data are combined to form a superquartet. (Similar approaches are possible applying distance or parsimony methods.) The resulting set of superquartets act as building blocks to construct an overall tree using an algorithm similar to quartet puzzling [SvH96] but guided by the above overlap graph. (This approach could also be used as a supertree method if the quartets are obtained from a set of genetrees.) By our medium level approach it is possible to include different evolutionary constraints acting on different genes. Thus, we do not treat the collection of different sequences as one big gene like when applying the total evidence approach naïvely. On the other hand, while the sequence information is not transmitted to the combination into an overall tree in the supertree approaches, the superquartets carry – at least locally – this information, that is subsequently included in the puzzling step. Hence, our method tries to bring together advantages from supertree and supermatrix approaches.

Our novel method is applied to the dataset of the grasses [Gra01, GPWG]. It exhibits a performance that is comparable to the best supertree and supermatrix methods (with regard to the GPWG classification) tested. We will discuss the results and some extensions.

References

[BE04] Olaf R. P. Bininda-Emonds, editor. *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*. Kluwer Academic, Dordrecht, The Netherlands, 2004.

[Fel81] Joseph Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[Gra01] Grass Phylogeny Working Group. Phylogeny and subfamilial classification of the grasses (poaceae). *Ann. Mo. Bot. Gard.*, 88:373–457, 2001.

[Klu89] Arnold G. Kluge. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Syst. Zool.*, 38:7–25, 1989.

[RRG01] Marc Robinson-Rechavi and Dan Graur. Usage optimization of unevenly sampled data through the combination of quartet trees: An eutherian draft phylogeny based on 640 nuclear and mitochondrial proteins. *Isr. J. Zool.*, 47:259–270, 2001.

[SvH96] Korbinian Strimmer and Arndt von Haeseler. Quartet puzzling: A quartet maximum–likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.*, 13:964–969, 1996.

Stochastic Models of Molecular Evolution: An Algebraical Analysis

Steffen Kläre

Heinrich-Heine-Universität Düsseldorf
Institut für Bioinformatik
klaere@cs.uni-duesseldorf.de

The main objective of phylogenetic reconstruction methods is the inference of an ancestral relationship between observed species. Generally, the relationship is visualized using so-called *evolutionary trees*.

From a stochastic point of view this task is formulated in the following way: Find a rooted tree $\mathcal{T}_\varrho = (\mathcal{V}, \mathcal{E}; \varrho)$ with vertex set $\mathcal{V}$, edge set $\mathcal{E}$ and root ϱ, and an associated stochastic process $\mathbf{X} : \mathcal{V} \to \mathcal{S}$, which assigns to every vertex $\alpha \in \mathcal{V}$ a state x_α from the genetical alphabet $\mathcal{S}$, such that the characterizing probability distribution of $\mathbf{X}$ best fits the observed data at the leaves.

Usually, the process $\mathbf{X}$ is said to have the *Markov property*, i.e. the process in a vertex α is given the process in its parent vertex $\mathrm{pa}(\alpha)$ conditionally independent of all its nondescendants. The Markov property is equivalent to the following equation (see e.g. [Lau99], sect. 1.5):

$$\mathrm{Prob}\Big(\bigcap_{\alpha\in\mathcal{V}} \{X_\alpha = x_\alpha\}\Big) = \mathrm{Prob}(X_\varrho = x_\varrho) \prod_{(\gamma,\beta)\in\mathcal{E}} \mathrm{Prob}(X_\beta = x_\beta | X_\gamma = x_\gamma). \tag{F}$$

With (F) the task of phylogenetic derivation can be formulated in the following way: Find a rooted tree $\mathcal{T}_\varrho$ with leaf set $\mathcal{L} = \{\beta_1, \ldots, \beta_n\}$ and an associated Markov process $\mathbf{X}$ such that the distribution given by

$$\mathrm{Prob}(X^{\mathcal{L}} = \underline{x}) = \sum_{\substack{\underline{y} \in \mathcal{S}^{n+m} \\ \underline{y}|_{\mathcal{L}} = \underline{x}}} \mathrm{Prob}(X_\varrho = y_\varrho) \prod_{(\alpha_1,\alpha_2)\in\mathcal{E}} \mathrm{Prob}(X_{\alpha_2} = y_{\alpha_2} | X_{\alpha_1} = y_{\alpha_1}) \tag{†}$$

best describes the data collected at the leaves. Apparently, equation (†) is a generalization of equation (2) in [Fel81], and thus also the basis for the Maximum Likelihood approach of phylogenetic inference.

The equation system (†) strongly depends on the structure of the rooted tree $\mathcal{T}_\varrho$. Since the particular tree structure is usually unknown, this is very unsatisfactory. Fortunately, [Cha96] showed that a Markov process on a tree can be reconstructed by its triple tree restrictions, i.e. by the restriction of $\mathbf{X}$ to every subtree consisting of three leaves and one inner vertex.

This result is used as a motivation to study the properties of the system (†) on triple trees. One general observation is, that (†) does not have a solution w.r.t. to every given leaf distribution. Hence, there must be a characterization of leaf distributions for which (†) has a solution. This characterization is a system of polynomials, each of which is called a *phylogenetic invariant* (see e.g. [AR03]).

We studied the model for three particular specifications, namely the general two state model, the Neyman N_k or general Jukes Cantor model, and the Kimura 2ST model. For all models phylogenetic invariants are inferred, and closed forms of algebraical solutions are derived. In addition, conditions for the existence of a Markov process to the given leaf distribution are computed. For the two state model an attempt is presented to extend the insights from triple trees to quartet trees.

References

[AR03] Elizabeth S. Allman and John A. Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Mathematical Biosciences*, 186(2):113–144, December 2003.

[Cha96] Joseph T. Chang. Full reconstruction of Markov models on Evolutionary Trees: Identifiability and consistency. *Mathematical Biosciences*, 137:51–73, 1996.

[Fel81] Joe Felsenstein. Evolutionary Trees from DNA sequences: A Maximum Likelihood approach. *Journal of Molecular Evolution*, 17(6):368–76, 1981.

[Lau99] Steffen L. Lauritzen. Causal inference from graphical models. Aalborg University, 1999.

Algorithmic & Technical Concepts in RAxML-V

Alexandros Stamatakis and Michael Ott

Institute of Computer Science, Foundation for Research and Technology-Hellas, P.O. Box 1385, GR-71110, Heraklion, Crete, Greece
Lehrstuhl für Rechnertechnik und Rechnerorganisation, Technische Universität München, Boltzmannstr. 3, D-85748, Garching b. München, Germany
stamatak@ics.forth.gr, ottmi@in.tum.de

The computation of ever larger as well as more accurate phylogenetic trees represents one of the grand challenges in High Performance Computing Bioinformatics. The size of trees which can be computed in reasonable time based on elaborate evolutionary models is limited by their severe computational cost. There exist two orthogonal research directions to solve this problem: *Firstly*, the development of novel, faster, and more accurate heuristic algorithms. *Secondly*, the application of high performance computing techniques.

The field has witnessed significant algorithmic advances over the last 2–3 years which allow for inference of large phylogenetic trees containing 500-1.000 sequences on a single PC processor within a couple of hours using maximum likelihood (ML). The main problem which high performance computing implementations of ML analyses face is that technical development drags behind algorithmic development, i.e. programs are parallelized that do not represent the state-of-the-art algorithms any more. We provide an overview of the general algorithmic *and* high performance computing concepts implemented in RAxML-V which are easily applicable to other tree building programs.

Algorithmic Concepts: Initially, we re-visit the Subtree Equality Vector (SEV) technique which accelerates the evaluation of the likelihood function by detecting equal alignment patterns in subtrees and re-using previously computed values. Depending on the CPU architecture, this method yields run time improvements between approximately $30 - 65\%$. Furthermore, the lazy subtree rearrangement technique is outlined which represents the key component of the hill-climbing heuristics implemented in RAxML-V [SLM05] and significantly improves upon tree quality. Finally, we address the new simulated annealing algorithm which has been implemented in RAxML-V.

Technical Concepts: Initially, we describe the MPI-based parallelization of RAxML which has been used to perform one of the largest ML-based phylogenetic analyses to date (comprising 10.000 taxa) on a Linux cluster. The main focus is on the new parallelization of RAxML-V for Symmetric Multi-Processing machines (SMPs) using OpenMP. The program achieves *significant* superlinear speedups for *long* alignments on 4-way AMD Opteron SMPs. The superlinear speedups for large alignments are caused by improved cache-efficiency and data locality. Moreover, the OpenMP implementation allows for inference of 500-taxon trees in less than 2 hours on a 4-way Opteron. Finally, initial results of the OpenMP parallelization of PHYML [GG03] are presented.

Current and Future Work: Current algorithmic work mainly focuses on the design and evaluation of divide-and-conquer algorithms. In cooperation with Usman Roshan we are currently evaluating the performance of using Rec-I-DCM3 [RMWW04] in conjunction with RAxML. Finally, we provide an overview over future High Performance Computing implementations of RAxML including a hybrid MPI/OpenMP parallelization and the exploitation of vector-like peripheral processors such as Graphics Processing Units (GPUs). The current RAxML-V source code as well as the respective publications are available for download at www.ics.forth.gr/~stamatak.

References

[GG03] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704, 2003.

[RMWW04] U. Roshan, B. M. E. Moret, T. Warnow, and T. L. Williams. Rec-i-dcm3: a fast algorithmic technique for reconstructing large phylogenetic trees. In *Proceedings of CSB2004*, Stanford, California, USA, 2004.

[SLM05] A. Stamatakis, T. Ludwig, and H. Meier. RAxML-III: A fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4):456–463, 2005.

An Evolutionary Space-Time Model with Varying Among Site Dependencies

Adi Stern and Tal Pupko

Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences
Tel Aviv University, Ramat Aviv 69978, Israel
{sternadi,talp}@post.tau.ac.il

Evolutionary models are a natural representation of the stochastic nature of evolution, allowing the introduction of realistic parameters in the model, within a robust statistical framework of inference. Better models are needed for accurate estimation of phylogenetic trees, inference of site-specific evolutionary rates and for ancestral sequence reconstruction.

Until recently, most models of evolution assumed that evolutionary rates at adjacent sites were independent. Yang [Yan95] and Felsenstein [FC96] suggested a model of evolution which takes into account a linear correlation between adjacent nucleotides by using a *Hidden Markov Model (HMM)*. We hereby refer to the model suggested by Yang [Yan95] as the D model. Yet such a model assumes that the level of correlation between rates at adjacent sites is equal at all sites of the protein. This assumption may be invalid: high correlation is expected in, for example, linear functional domains. On the other hand when we consider the 3-dimensional structure of the protein, low correlation is expected, since the interaction between distant sites imposes independence of rates in the linear sequences (e.g., active sites).

We have developed a model which allows the level of correlation between rates at adjacent sites to vary. This is done by using an extended HMM, which takes into account both auto-correlation at certain regions of the protein as well as rate independence at other regions. We extend the hidden states of the Markov chain to incorporate these two types of situations. Thus, the number of possible hidden states is doubled to include two different sets: R_D represents the set of rates given dependence and R_I represents the set of rates given independence. The set of values for the hidden states is now $R_D \bigcup R_I$. Two parameters, λ_1 and λ_2 represented the probability of a transition between R_D and R_I and vice-versa, respectively. We refer to this novel model as $D + I$, indicating dependence and independence.

In order to test the utility of the new model, ten datasets were selected from the datasets published by Aloy [AQAS01]. The likelihood ratio test (LRT) was used in order to test whether the $D + I$ model fitted a particular dataset significantly better than the D model. Our results, shown in Table 1, indicate that this was indeed true in all examined cases.

PDB ID	D	$D+I$	PDB ID	D	$D+I$
1bro	-38770.5	-38709.4	1bls	-23460.63	-23437.99
1orq	-19694.2	-19663.33	1clx	-28361.59	-28335.03
1dxy	-35306.99	-35254.32	1a0p	-16428.67	-16415.19
1bpl	-24631.4	-24608.5	1ast	-19895.88	-19869.57
1aih	-7610.76	-7595.25	1ay1	-9964.68	-9944.93

Table 1: Maximum log-likelihood for the analysis of 10 datasets under the D model and the $D+I$ model. The datasets are referred to by their protein data bank (PDB) [BWF$^+$00] identifiers. In all datasets, the LRT between the D model and the $D + I$ model is significant (P value < 0.001)

References

[AQAS01] P. Aloy, E. Querol, F.X. Aviles, and M.J. Sternberg. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J Mol Biol*, 311(2):395–485, 2001.

[BWF$^+$00] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[FC96] J. Felsenstein and G.A. Churchill. A hidden markov model approach to variation among sites in rate evolution. *Mol Biol Evol*, 13(1):93–104, 1996.

[Yan95] Z. Yang. A space-time model for the evolution of dna sequences. *Genetics*, 139:993–1005, 1995.

Evolution of plant telomeres – step back or forward?

E. Sykorova[1,2], K. Neplechova[1], M. Sklenickova[2], K.Y. Lim[3], F. Blattner[4], M. W. Chase[5], A. R. Leitch[3] and J. Fajkus[1,2]

[1]Institute of Biophysics, Czech Academy of Sciences, Kralovopolska 135, 61265 Brno, Czech Republic,
[2] Department of Functional Genomics and Proteomics, Masaryk University, Kotlarska 2, 61137 Brno, Czech Republic
[3] School of Biological Sciences, Queen Mary University of London, Mile End Road, E1 4NS London, UK
[4] Department of Taxonomy, Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany
[5] Jodrell Laboratory, Royal Botanic Gardens, Kew, TW9 3AB Richmond, UK

The search for missing typical plant telomeres in Asparagales revealed that this monocotyledonous plant order can be subdivided into three groups of plants differing substantially in the type of their telomeres - plant-type (e.g. orchids), human-type (e.g. asparagus), unknown type (e.g. onion). The presence of different types of telomere corresponds with phylogeny and it defines two evolutionary points where the change had happened [SLK03+]. The first evolutionary point divides the families with the plant and the human type of telomeres with telomerases maintaining corresponding telomeric motifs, the second point divides the families with the human and the unknown type of telomeres (Alliaceae). To further define this switch point we started screening the Alliaceae family to find out where the loss of human type of telomeres had happened. Our current results suggest that the "big change" connected with the change of telomere type separates *Allium* genus from other members of Alliaceae. This finding moves us closer to a solution of the mystery of onion telomeres and their maintenance.

Although the previously suggested models of function of alternative onion telomeres based on satellite repeats or transposable elements could explain the mechanism of replenishment of chromosome end sequences, the other telomeric functions (e.g., the end capping) would require substantial adaptation of the protein machinery involved. Nowadays, missing plant telomeres are not exceptional. The similar evolution event occurs in Solanaceae where model plants (tobacco, tomato) have typical plant-type of telomeres but these telomeres were lost in genus *Cestrum* [SLC03+, SLFL0 3]. However, evolution of telomeres seems to be more complicated in insects and arthropods [SMT99, VKT05+], the loss of typical telomeres was reported recently in beetles and spiders. It raises general questions about evolution of telomeres, mechanisms of maintenance and adaptation of telomere-binding proteins [RSD04+] in response to such evolutionary event.

This work was supported by GACR (204/04/P105, 521/05/0055), GA AS CR (A600040505) and the institutional support (MSM0021622415, AV0Z50040507).

References:

[SLK03+] E. Sykorova, K. Lim, Z. Kunicka, M. Chase, M. Bennett, J. Fajkus, A. Leitch. Telomere variability in the monocotyledonous plant order Asparagales. (2003) Proc R Soc Lond B Biol Sci. 270(1527): 1893-904.

[SLC03+] E. Sykorova, K. Lim, M. Chase, S. Knapp, I. Leitch, A. Leitch, J. Fajkus. The absence of Arabidopsis-type telomeres in Cestrum and closely related genera Vestia and Sessea (Solanaceae): first evidence from eudicots. (2003) Plant J. 34(3):283-91.

[SLFL03] E. Sykorova, K. Lim, J. Fajkus, A. Leitch. The signature of the Cestrum genome suggests an evolutionary response to the loss of (TTTAGGG)n telomeres. (2003) Chromosoma 112(4):164-72.

[RSD04+] G. Rotkova, M. Sklenickova, M. Dvorackova, E. Sykorova, A. Leitch, J. Fajkus. An evolutionary change in telomere sequence motif within the plant section Asparagales had significance for telomere nucleoprotein complexes. (2004) Cytogenet Genome Res. 107(1-2):132-8.

[SMT99] K. Sahara, F. Marec, W. Traut. TTAGG telomeric repeats in chromosomes of some insects and other arthropods. (1999) Chromosome Res. 7(6):449-60.

[VKT05+] M. Vitkova, J. Kral, W. Traut, J. Zrzavy, F. Marec. The evolutionary origin of insect telomeric repeats, (TTAGG)n. (2005) Chromosome Res. 13(3):145-56.