UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 226

Phylogenetic Support and Chloroplast Genome Evolution in *Sileneae* (Caryophyllaceae)

PER ERIXON

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2006

ISSN 1651-6214
ISBN 91-554-6668-0
urn:nbn:se:uu:diva-7166

Dissertation presented at Uppsala University to be publicly examined in Zootis-salen, EBC, Villavägen 9, Uppsala, Friday, October 20, 2006 at 09:00 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

Erixon, P. 2006. Phylogenetic Support and Chloroplast Genome Evolution in *Sileneae* (Caryophyllaceae). Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 226. 41 pp. Uppsala. ISBN 91-554-6668-0.

Evolutionary biology is dependent on accurate phylogenies. In this thesis two branch support methods, Bayesian posterior probablities and bootstrap frequencies, were evaluated with simulated data and empirical data from the chloroplast genome. Bayesian inference was found to be more powerful and less conservative than maximum likelihood bootstrapping, but considerably more sensitive to choice of parameters. Bayesian inference increased in power when data were underparameterized, but the associated increase in type I error was comparatively larger.

The chloroplast DNA phylogeny of the tribe *Sileneae* (Caryophyllaceae) was inferred by analysis of 33,149 aligned nucleotide bases representing 24 taxa. The position of the SW Anatolian taxa *Silene cryptoneura* and *S. sordida* strongly disagreed with previous studies on nuclear DNA sequence data, and indicate a possible case of homoploid hybrid origin. *Silene atocioides* and *S. aegyptiaca* formed a sister group to *Lychnis* and remaining *Silene*, thus suggesting that *Silene* may be paraphyletic, despite recent revisions based on molecular data. Several nodes in the phylogeny remained poorly supported, despite large amounts of data. Additional sequence sampling is not expected to solve this problem. The main reason for poor resolution is probably a combination of rapid radiation and substitution rate hererogeneity. Apparent incongruent patterns between different regions of the chloroplast genome are evaluated with ancient interspecific chloroplast recombination as explanatory model.

Extremely elevated substitution rates in the exons of the plastid *clpP* gene was documented in *Oenothera* and three separate lineages of *Sileneae*. Introns have been lost in some of the lineages, but where present, intron sequences have a markedly slower substitution rate, similar to the rates found in other introns of their genomes. Three branches in the phylogeny show significant whole gene positive selection. In two of the lineages multiple partial copies of the gene were found.

Keywords: Phylogenetics, Bayesian inference, Bootstrapping, cpDNA, Sileneae, Interspecific chloroplast recombination, Hybridization, clpP, Positive selection, Extreme substitution rates

Per Erixon, Department of Evolution, Genomics and Systematics, Systematic Botany, Norbyv. 18D, Uppsala University, SE-75236 Uppsala, Sweden

ISSN 1651-6214
ISBN 91-554-6668-0
urn:nbn:se:uu:diva-7166 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-7166)

"If it weren't for reality his predictions would have come true."

"Always do what you're told and you'll never get lost, nor find your way."

Dartwill Aquila

Cover: *Lychnis chalcedonica* L. in Uppsala Botanical Garden.
Bayesian phylogram of eudicot *clpP1* gene sequences.

List of Papers

This thesis is based on the following papers, which are referred to by their Roman numerals:

I **Erixon, P**., B. Svennblad, T. Britton, and B. Oxelman. 2003. Reliability of Bayesian Posterior Probabilities and Bootstrap Frequencies in Phylogenetics. Systematic Biology 52: 665-673.

II **Erixon, P**., and B. Oxelman. The effect of under-parameterization and character dependency on Bayesian posterior probabilities and maximum likelihood bootstrap frequencies on empirical data. Manuscript.

III **Erixon, P**., and B. Oxelman. Reticulate or treelike chloroplast DNA evolution in *Sileneae* (Caryophyllaceae)? Manuscript.

IV **Erixon, P**., and B. Oxelman. Elevated substitution rates, putative positive selection, duplications, and intron evolution in the plastid *clpP* gene in *Oenothera* and *Sileneae.* Manuscript.

Paper I is reprinted with the publisher's kind permission.

All papers included in this thesis are written by the first author, with comments and suggestions given by the co-authors. The studies were planned in cooperation with the co-authors. The first author is responsible for all analyses and laboratory work, but the statistical analysis for figure 2 in paper I was conducted by the second author of that paper.

Contents

Introduction

Phylogenetics is the study of evolutionary relatedness among various groups of organisms. In theory, the speciation process results in a bifurcating genealogy, a phylogeny. In molecular phylogenetics the means for reconstructing this phylogeny are usually DNA sequences from extant species and a particular inference method. Phylogenetics is a historical science, but because the evolutionary process spans over such large time-scales, it can rarely be directly observed, we have to infer it from the pattern manifested in present-day organisms. Before the age of molecular biology, systematists primarily used morphological similarities to infer relationships. For many organismal groups, the number of potential informative morphological characters is small, and because morphology sometimes reflects adaption to specific conditions, the associated uncertainty can be large.

Data in angiosperm molecular systematics

The characters used in molecular phylogenetics are usually the individual sites in an aligned matrix of DNA sequences. Morphological characters have to be interpreted to infer homology, but for a specific nucleotide base in a DNA sequence it is impossible to discriminate between homology and non-homology, an A that has changed to a G and then back to an A, is identical to a non-changed A. The homology assessment is guided by the overall similarity in the alignment of the DNA sequences. Because sequences from the same region, but for different taxa, often are unequal in length, gaps have to be inserted to the sequences. The homology assessments in the alignment process become more difficult with decreased sequence similarity.

The amount of sequence data is almost inexhaustible in eukaryotic organisms, e.g. the genomes of *Silene latifolia* (Fig. 6) contain 2.7 billion base pairs (bp) (Bennett & Leitch, 1997). Most phylogenetic studies encompass less than one millionth of that. An overwhelming majority of the genetic material is found in the nuclear genome, only 0.15 million bp are found in the chloroplast genome (Spinach, *Spinacia oleracea*) and 0.37 million bp in the mitochondrion (Beet, *Beta vulgaris*).

The knowledge of angiosperm phylogeny has increased substantially during the last decade, as manifested in the papers by the Angiosperm Phylogeny Group (APG, 1998; APG II, 2003) for higher level systematics, as well as the vast literature on various groups at lower levels. A majority of the DNA sequence data used in those studies comes from the chloroplast genome (cpDNA), and a restricted set of regions. A survey of Shaw et al. (2005) of papers published from 1995 through 2002 in *American Journal of Botany*, *Systematic Botany*, *Molecular Phylogenetics and Evolution*, and *Plant Systematics and Evolution* found that the utilization of non-coding cpDNA has increased during recent years, but that 77% of the surveyed studies used some portion of either *trnK*-*matK*-*trnK*, the *trnL* intron, and/or the *trnL*-*trnF* spacer.

In addition to cpDNA, many studies are based on nuclear ribosomal DNA (nrDNA), such as the internal transcribed spacers (ITS) or 18S. These regions exist in hundreds of copies in tandem arrays in the nuclear genome and because concerted evolution tends to homogenize the copies, they are usually treated as a single locus. The popularity of cpDNA and nrDNA data is mainly due to practical reasons. Many researchers have used them, there are a wide range of PCR primers, and there is a large amount of reference sequences on GenBank. Many cpDNA genomes of flowering plants have also been completely sequenced (http://chloroplast.cbio.psu.edu/). Because of the hundreds of copies in each cell, sequences from cpDNA and nrDNA are often easily amplified. The use of real nuclear genes in phylogenetics is much more limited, despite the almost inexhaustible amount of data they represent.

Currently, systematists often use DNA sequences only as carriers of historical information. Because DNA accumulates mutations over time, and those changes (if germinal) are vertically transferred through generations, a shared derived character state (A, C, G, or T) is evidence of common ancestry. In most cases a very limited amount of DNA is sequenced to infer the phylogenetic relationships of an organismal group. The choice of sequence region is guided by practical reasons, as discussed above, but also on the particular phylogenetic problem. There is a common conception that lower level systematics should use fast evolving regions, whereas higher level systematics should use slow evolving regions. This is intuitive, at first, because closely related taxa are temporally less separated, so in order to observe sufficient changes (differences in sequences from different taxa), the sequences cannot be too slow evolving. For example, the complete chloroplast genomes of *Nicotiana tabacum* (155,943 bp) and *Nicotiana sylvestris* (155,941 bp) only differ in 7 (!) sites (Yukawa et al., 2006). For distantly related taxa high variability can cause uncertainty in sequence alignment, but the main concern is generally multiple saturating substitution (e.g. Goremykin et al., 2003).

But what is the reason for different substitution rates in coding and non-coding DNA regions or between different sites in coding regions? No DNA sequences are completely unconstrained, but most sites in most non-coding regions of DNA are probably fairly unconstrained (but see Andolfatto, 2005). This means that the probability of observing a change in one site is not very different from any other site, i.e. neutral mutations result in random substitutions. When inferring phylogenetic relationships, this is advantageous, because modelling of the evolutionary process becomes less complex. Most genes, on the other hand, are under negative selection. The most critical changes are those that alter the amino acid composition, i.e. non-synonymous substitutions. Synonymous substitutions are more similar to substitutions in non-coding DNA, because they have generally no direct impact on the function of the protein, and thus little effect on the organism's fitness. The difference in actual mutation rate between sites in the same gene is minimal, and the difference in observed substitution rate arises because most of the non-synonymous *mutations* are eliminated from the population by purifying selection (Grauer and Li, 2000). The use of more slowly evolving protein coding genes for distantly related taxa is not straight forward, because the within gene rate heterogeneity is often considerably higher. One common way to account for this is to exclude the third codon position from the analysis (e.g. Goremykin et al., 2003), or parameterize the codon positions differently (e.g. Brandley et al., 2005), following the idea that synonymous substitutions are more common, and occur more frequently in third positions. This instrument is, however, rather blunt, because not all third position substitutions are synonymous and one can also easily imagine amino acids that are almost freely interchangeable, and synonymous substitution destructing important protein secondary structures. The overall evolutionary rate is a poor indication of phylogenetic utility, the mutational pattern of individual sites is more important (Müller et al., 2006).

Adaptive or positive selection, i.e. when the observed rate of non-synonymous substitutions are higher than synonymous in specific sites can occur when amino acid change is highly advantageous, because such mutations undergo fixation in a population much more rapidly than neutral mutations (Grauer and Li, 2000). Even if individual sites are under positive selection, non-synonymous substitutions rarely dominate when the entire gene is considered.

Phylogenetic support

Accurate phylogenies are important, because they provide a framework for addressing a wide range of biological questions, of which classification is just one. To understand morphological character evolution, biogeographical patterns, the frequency and mode of molecular evolutionary processes, the timing of phylogenetic events, and more, phylogenetic trees are essential.

How to find the best tree topology? Traditionally, the basic approach has been to evaluate *all* possible topologies and chose *the* best topology according to an optimality criterion, i.e. chose the topology that minimizes evolutionary change (maximum parsimony) or the topology, with branch lengths, that maximize the probability of observing the data given a specific evolutionary model (maximum likelihood).

During the last two decades the focus of phylogenetic research has shifted from finding the best topology by maximizing an optimality criterion (e.g. parsimony or maximum likelihood) or applying a particular algorithm (e.g. neighbor-joining) to quantification of uncertainty in the data by evaluating the support for certain clades. Bootstrap values (Felsenstein, 1985), Bremer support values (Bremer, 1988, 1994), jackknife values (Farris et al., 1996), and Bayesian posterior probabilities (Rannala and Yang, 1996; Yang and Rannala, 1997; Huelsenbeck et al., 2001) are commonly used measures.

An ideal support value for a clade could represent the probability that the depicted relationship is true. However, it is unrealistic to have such expectations. Another interpretation of clade support could be the probability that we would infer the same tree if we gather other data of the same size. Put in this way support for a clade would be intuitively appealing, because it is reasonable to think that a phylogeny based on small amounts of data is more likely to be inaccurate than a phylogeny based on large amounts of data. Support measures based on very small amounts of data are indeed generally low and resolution in phylogenetic trees is often improved if additional data are gathered for the same set of taxa. The probability that a specific clade would persist when more data are gathered is unfortunately not only dependent on sampling error, it is also dependent, and heavily so, on error due to violations of assumptions. To illustrate this point, consider this example: Suppose that you want to estimate the frequency of different colors of *otherwise identical* marbles well mixed in a large sack. The error of your estimate based on drawing random marbles from the sack is only dependent on the sample size, i.e. the accuracy is better if you estimate the frequencies on the basis of 10000 sampled marbles compared to only 100. The situation in phylogenetics is very different. To begin with sequences are rarely an unbiased sample, and the sites in the sequence are not independently and identically distributed (i.i.d.). These violated assumptions affect any method of phylogenetic inference. On top of this are all assumptions violated in the specific phylogenetic inference method used. Inefficiency in a method can be remedied by additional sequencing, but all methods are inconsistent under some circumstances, and inconsistency does not benefit from additional sequencing.

What is Bayesian inference of phylogeny?

Some years ago there were many heated discussions about the pros and cons of maximum parsimony or maximum likelihood as optimality criterion. Proponents of parsimony emphasized the intuitive simplicity (Ockham's razor), whereas likelihoodists emphasized that parsimony is inconsistent under certain conditions (Felsenstein, 1978). To most systematists, the choice of method probably had more practical arguments; many used parsimony because with a growing number of taxa to analyse, inference by maximum likelihood was just not feasible. Problems with unequal substitution rates and long branches were ignored, or, when possible, minimized by denser taxon sampling.

In 1996, there were proposals of a new method for phylogenetic inference (Rannala & Yang, 1996; Mau, 1996; Li, 1996). The method had a completely different statistical framework, but used the familiar likelihood function to evaluate tree topologies. Because most systematists are dependent on software implementations, this new method did not hit it off until four years later with the first release of MrBayes, where the computational burden inherent in Bayesian statistics was circumvented by the application of Markov chain Monte Carlo methods (Huelsenbeck & Ronquist, 2001). The impact of Bayesian inference on phylogenetics has since been huge. The principle publication of the software (Huelsenbeck & Ronquist, 2001) has in only a few years received 2122 citations (ISI Web of Science, 8th September 2006). By comparison the initial publication of Rannala and Yang (1996) has received only a tenth of that in twice as long time period.

The almost immediate success of Bayesian phylogenetic inference, had hardly to do with the conceptual differences relative to maximum likelihood or parsimony, but rather to its similarities with maximum likelihood. In phylogenetics, computational speed is a big issue! Suddenly, model based inference with relatively complex substitution models could be performed on data sets with more than a handful of taxa. Was the hegemony of parsimony finally over? The front figures of the new method were more than willing to boost this first impression: “[Bayesian inference] is roughly equivalent to performing a maximum likelihood analysis with bootstrap resampling, but much faster” (Huelsenbeck et al., 2001). However, the support values from the Bayesian analyses, the posterior probabilities, were often found to be considerably higher than, e.g. frequencies from maximum likelihood or parsimony bootstrap analyses. Some authors take the cynical standpoint that part of the success of Bayesian inference is due to the higher support values per se (Randle et al., 2005). Evolutionary hypotheses are more exciting if they are based on a fully resolved and strongly supported tree.

The main advantage, beyond practical issues, of Bayesian phylogenetic inference is that the posterior distributions of many parameters, including tree topology, can be inferred simultaneously. To estimate, for example, the transition/transversion ratio of a particular gene, there is no need to rely on a single point estimate of the tree topology.

Another appealing feature of Bayesian phylogenetic inference is that it gives probabilities for our hypotheses of interest, the phylogenies. Likelihoods have a solid statistical framework and they are very useful, but difficult to interpret, because they represent the probability of the data given the hypothesis, rather than the more comprehensible probability of the hypothesis given the data. To say something of the probability of the hypothesis, maximum likelihood has to rely external methods, such as bootstrapping.

In order for Bayesian inference to be able to calculate posterior probabilities for phylogenies, prior distributions of all parameters have to be specified without reference to the data. This subjectivity is a major criticism of Bayesian statistics in general, not only to its application in phylogenetics. The priors always influence the posterior, but given reasonable amount of data the effect of the prior on the conclusions are expected to be small (but see below).

Finally, an important difference between Bayesian inference and maximum likelihood is that the former bases its estimates on the marginal likelihood, whereas the latter uses profile likelihood. If the ML-topology has a very restricted range of parameter values, the posterior probability can favor a different topology, also without the influence of the prior. However, such a restricted ML-topology would probably receive low bootstrap support. For an excellent review of the underlying principles of Bayesian inference and maximum likelihood see Holder and Lewis (2003).

Aims

In paper I we used a simulation approach to test the hypothesis that there is no difference between maximum likelihood bootstrap frequencies and Bayesian posterior probabilities. We also investigated the error rate and power of the two methods under both the correct substitution model and with under-parameterization.

In paper II we wanted to explore if under-parameterization and character dependency (another potential cause of model misspecification) in empirical data could explain some of the observed difference between the two methods.

In paper III we used a large cpDNA data set to infer relationships in *Sileneae* (Caryophyllaceae). We wanted to investigate if all branches in the phylogeny could be confidently resolved given enough data. If not, we wanted to explore possible causes of weak support.

As a result of paper III, we found that the *clpP1* gene in some taxa of *Sileneae* was rapidly evolving. In paper IV we investigated the phylogeny of the *clpP* gene family in flowering plants and explored how gene duplication, elevated substitution rates, positive selection, and structural changes correlate.

Paper I

In paper I we evaluated the relative errors in support values, when interpreted as probabilities, received from Bayesian inference (BAYES) and maximum likelihood bootstrapping (MLBOOT). Several previous studies (e.g., Karol et al., 2001; Murphy et al., 2001; Leaché and Reeder, 2002; Whittingham et al., 2002) had observed that Bayesian posterior probabilities were generally higher then MLBOOT frequencies in comparisons on empirical data. Is the apparent strong power of Bayesian inference associated with a cost in the form of more frequent false positives?

Efron et al. (1996) claimed that bootstrap frequencies can be interpreted as posterior probabilities under certain circumstances. Those sentences have been frequently cited (e.g., Larget and Simon, 1999; Cummings et al., 2003; Simmons et al., 2004), and at least implicitly with claims that bootstrap support values and Bayesian posterior probabilities should be approximately equal. Because the theory did not seem to harmonize with the empirical findings, we wanted to investigate if there was a difference between the two methods under identical model assumptions.

Our simulations showed that Bayesian posterior probabilities, on average, are higher than MLBOOT frequencies for well-supported clades. BAYES has a larger proportion of high (>95%) and low (<25%) support values for true clades compared to MLBOOT. The pattern is strongly accentuated with under-parameterization, i.e. when data generated under a complex model are analyzed with a model with fewer parameters. There are almost twice as many true clades with support >95% when BAYES uses a "too simple" model (Fig. 1).

With large amounts of data and under the correct substitution model the difference between the two models is small, but significant. Our results indicate that both BAYES and MLBOOT are conservative under the correct model, if interpreted as in traditional frequency statistics (but see Huelsenbeck and Rannala, 2004). MLBOOT is, however, more conservative.

In our simulations bootstrap values *above* 70% were correct 95% of the times, an observation also made by Hillis and Bull (1993). That paper is often sited as support for using 70% bootstrap support as cut-off for reliable clade. It is important to note, however, that a bootstrap value of 70% was found false 16% of the times (Paper I, Fig. 2).

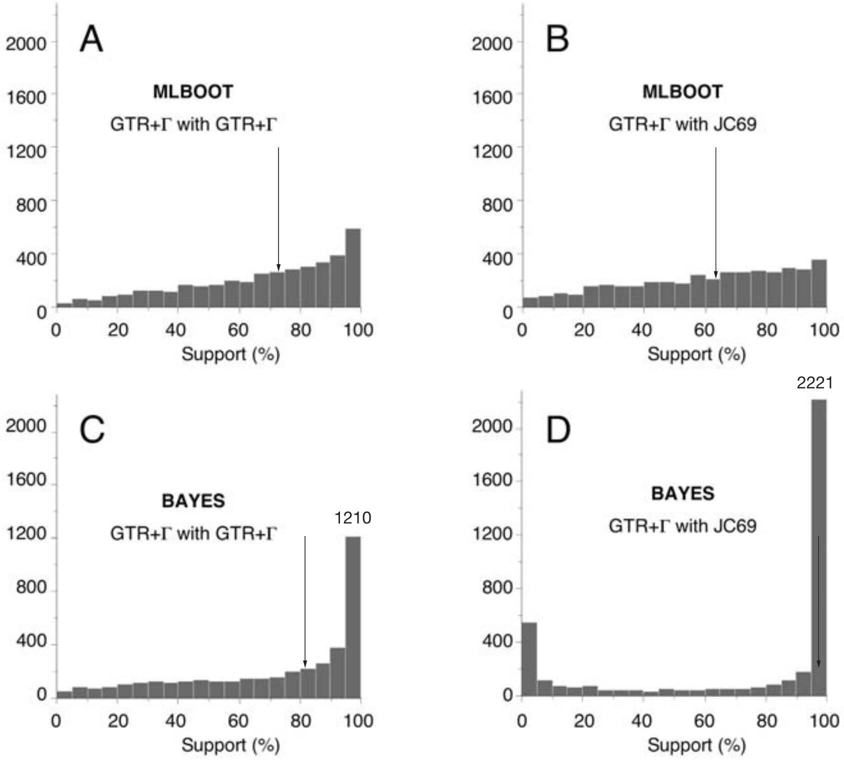

Figure 1. Histograms of support value distributions for true clades (simulated under the GTR+ Γ model). The y-axis represents number of true clades in each support category. Total number of support values for true clades is 4,000. (a) MLBOOT analyzed with GTR+Γ; (b) MLBOOT analyzed with JC69; (c) BAYES analyzed with GTR+ Γ; (d) BAYES analyzed with JC69. Arrows indicate median support.

A short review of more recent results

Many studies that have evaluated the difference between Bayesian posterior probabilities and bootstrap frequencies conclude that the latter, on average, are lower for the best supported clades. In the extensive comparison of Bayesian posterior probabilities and bootstrap frequencies in the four-taxon case under 1369 different branch length combination done by Cummings et al. (2003), the conclusion was that different tree topologies (i.e. branch lengths) shift the bias between the two measures.

The relationship between Bayesian posterior probabilities and bootstrap frequencies is not a straight line, but rather a sigmoidal curve (Paper I, Fig. 2). This is intuitive because if high support values are higher for Bayesian inference, low support values must be lower for Bayesian inference, because the frequencies sum to one per internal branch of the tree. If there are three possible topologies and the best topology A gets a posterior probability of 80% and a bootstrap frequency of 70%, the bootstrap frequencies for at least one of the topologies B or C are higher than the posterior probabilities, because they sum to 30% (1-0.70), while the posterior probabilities only sum to 20% (1-0.80). It is thus trivial to conclude that bootstrap frequencies are higher than posterior probabilities for correct topologies in situations when both methods put most of the support on incorrect topologies.

The "break" point in the sigmoidal curve, i.e. where the bootstrap frequencies change from being higher than the posterior probabilities to being lower, was by Cummings et al. (2003) concluded to be at support values around 93% (cf. Paper I, Fig. 2; <50%). A problem with this conclusion is that it is not based on the comparison of the individual data sets, but only on the average support values from each tree topology in the simulations. This can be misleading when support distributions are not identical. All simulations and empirical studies we know of, except that of Cummings et al. (2003) discussed above, have found that Bayesian posterior probabilities generally are higher than maximum likelihood bootstrap frequencies for well-supported groups. This led us to scrutinize the statement by Efron et al. (1996) to fully understand its meaning. In Svennblad et al. (2006) we show that the statement by Efron et al. (1996) is true under the assumptions they give, but that, in general phylogenetic inference, those assumptions are violated. The uniform prior density on data patterns (p) necessary for Bayesian posterior probabilities to be approximately equal to maximum likelihood bootstrap frequencies is not a parameter used in current implementations of Bayesian phylogenetic inference. In Svennblad et al. (2006), we also show, analytically, that some individual data patterns, e.g. XXYZ (two taxa share a state and the other two have different states), are separately informative in Bayesian inference, but not in Maximum likelihood. The explanatory power of this finding for the observed difference between Bayesian posterior probabilities and ML bootstrap frequencies remains, however, to be investigated.

Huelsenbeck and Rannala (2004) pointed out that a shortcoming of all previous simulation studies comparing Bayesian posterior probabilities and maximum likelihood bootstrap frequencies, including ours (Paper I), is that priors are ignored in the simulations and that parameters are treated as fixed, rather than as variables. In their own simulation (Huelsenbeck & Rannala, 2004), correcting for this flaw, they found a good correlation (although this is not statistically quantified) between posterior probability and probability of being correct, when the simulating and analysing models were identical.

The posterior probability of a phylogenetic tree is the probability that the tree is correct, assuming that the model is correct and all assumptions are met, but as Huelsenbeck and Rannala (2004) put it: "...all bets are off when the assumptions of a Bayesian analysis are not satisfied". Several studies have shown that Bayesian inference can assign excessively high support to incorrect topologies, when the data are analyzed with an oversimplified model (e.g. Buckley 2002; Suzuki et al. 2002; Lemmon & Moriarty 2004; Nylander et al. 2004). In our simulation, when data simulated under the GTR+Γ model were analyzed with the Jukes-Cantor model the Bayesian posterior probabilities were also found to be excessively high. The error rate of support values >95% was almost 16% (Paper I, Table 2). Analysis of empirical data probably always suffer from model misspecification, in some sense, even if the model that best fit the data, of the available ones, is used.

The most fundamental conceptual difference between Bayesian Inference and maximum likelihood is that prior probability distributions for all parameters in an analysis, including topology and branch lengths, have to be specified in Bayesian inference. A prior probability is the probability of a parameter before the data have been observed. It can be considered an advantage to incorporate prior beliefs in the analysis, but in systematics this is rarely done. Often, the researcher do not want any hypothesis to be favored a priori, but rather have the data dominate the outcome of the analysis. In theory, this can be achieved by making the prior distribution uninformative, i.e. to make any hypothesis equally probable. If, for example, the topology parameter is considered, an uninformative prior would put the same prior probability on all possible topologies. However, the choice of uninformative or flat priors is a very complex problem. On unbounded quantities, such as branch lengths, it is impossible, and probably not even desirable, to give all possible outcomes equal prior probability (Felsenstein, 2004). Even the seemingly clear-cut case of topology priors was shown to be problematic by Pickett and Randle (2005), because given an uniform topology prior, the clade priors cannot be uniform in unrooted trees with more than five taxa. The importance of the effect of unequal clade priors is, however, unclear (Picket & Randle, 2005; Brandley et al., 2006; Randle & Pickett, 2006).

If truly uninformative priors are unattainable, it is important that the signal in the data is sufficient in order for the likelihood function to overwhelm the influence of priors. Because of the complexity of empirical data it is difficult to assess when this is accomplished. In analyses containing many poorly supported bipartitions or in which bipartition posteriors are strongly dependent on the estimated model, different model priors can probably affect topology estimates significantly (Zwickl & Holder, 2004).

Recent developments have made it possible to do Bayesian inference with different parameterization for different partitions, and also combine morphological and sequence data in the same analysis (Nylander et al., 2004). Previously, combined analysis of morphological and sequence data was only possible with parsimony. Current implementations of maximum likelihood do not have this option. The number of parameters to estimate in the analysis rapidly grows, when the data are partitioned, and as a consequence the amount of data per substitution parameter decreases. This results in increased variance in the parameter estimates. More partitions, i.e. less data per parameter, also result in a stronger effect of the prior distribution on the posterior distribution. Several studies have found that increased partitioning results in dramatically improved likelihood scores (e.g. Nylander et al., 2004; Brandley et al., 2005; Strugnell et al., 2005). Although simulation studies have shown that over-parameterization generally is less problematic than under-parameterization for Bayesian inference (Paper I; Huelsenbeck & Rannala, 2004), excessive parameterization of empirical data is less investigated. Smedmark et al. (2006) experienced insurmountable convergence problems when partitions were differently parameterized. Despite extensive alteration of the MCMC proposal parameters, they could not get reproducible results with mixed models (Smedmark et al., 2006).

The computational efficiency of Bayesian phylogenetic inference using MCMC allow for more parameter-rich, and thus realistic substitution models. Development of such models is maybe more promising than excessive partitioning. Any *a priori* partitioning of the data is arbitrary and the problem of large sampling error in small partitions should not be ignored.

The gamma parameter, accounting for substitution rate heterogeneity, is often found to be the most important parameter in the substitution model (Nylander et al., 2004; Huelsenbeck & Rannala, 2004), but it assumes temporally constant rates for each site. The rate in a specific site probably changes during the course of evolution, and in different parts of the tree. The covarion model allows sites to be turned on and off on different branches in the tree. When it is off, no substitutions are possible and when it is on, substitutions occur according to the specified substitution model. Few studies have employed the covarion model, but at least for some data it can strongly improve the likelihood score (Shalchian-Tabrizi et al., 2006; Smedmark et al., 2006). Smedmark et al. (2006) actually found the covarion parameter to be more important than the gamma parameter for their data.

Paper II

The difference between Bayesian posterior probabilities and maximum likelihood under the correct model of evolution and large amount of data was small in our simulation study, around one %-unit difference in the support range 80-90% (Paper I, Fig. 4 and Fig. 5). Comparisons based on empirical data generally show more substantial differences (e.g. Douady et al., 2003; Leaché & Reeder, 2002; Karol et al., 2001). One reason for this discrepancy can be that the two methods behave differently under model misspecification and/or shortage of data.

An advantage of simulation-based comparisons is that the true underlying tree and model of sequence evolution are known. This gives the possibility to study the behavior of different methods both under the correct model of sequence evolution and under model misspecification. An important disadvantage is, however, that problems posed by the complex process of real evolution cannot be investigated.

In paper II we wanted to explore if under-parameterization and character dependency (another potential cause of model misspecification) in empirical data could explain some of the observed difference between the two methods. A data set of 25 kb cpDNA sequence was partitioned in non-overlapping sets of 1000 and 2500 bp. This was done both for blocks of adjacent characters and characters equidistantly sampled. We found no significant effect of character dependency in our data, and could thus not evaluate its relative effects on the two inference methods.

The maximum likelihood for the complete sequence matrix under the best approximating model available (GTR+iΓ) was compared with the unconstrained likelihood calculated under a multinomial model. The ML tree plus the GTR+iΓ model was rejected, indicating that the "best" model fit the data poorly.

The data partitions were analysed both under the JC69 and the GTR+iΓ substitution model and the behavior of Bayesian inference and maximum likelihood bootstrapping was evaluated by comparing the sums of support values in each partition. Support sums were calculated both for all resolved clades (unfixed topology) and the clades specific to the total evidence tree (fixed topology).

Maximum likelihood bootstrapping was found to be insensitive to the parameterization and very few partitions contained support for nodes not found in the total evidence analysis (Fig. 2). In contrast, the support sums in the Bayesian analyses increased with stronger under-parameterization (JC69) and also with the unfixed topology (Fig. 2), i.e. several nodes not in the total evidence topology receive support in the individual data partitions. The significant interaction between the substitution model and the topology factor (fixed/unfixed) for the Bayesian method indicate, under the assumption that the total evidence topology is true, that the increase in posterior probabilities with a too simplified model is mostly caused by false positives.

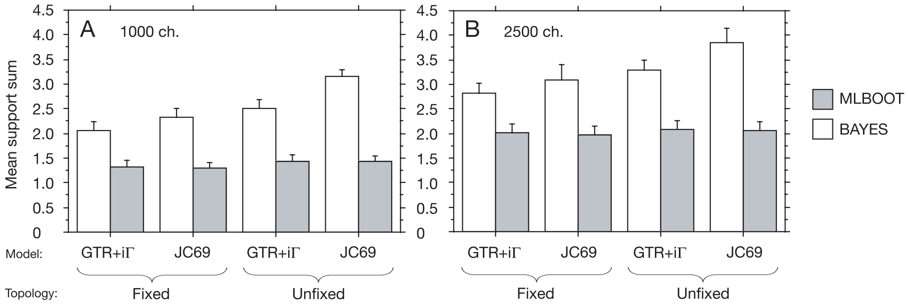

Figure 2. Interaction bar chart for factors: Method (BAYES/MLBOOT), Model (JC69/GTR+iΓ), and Topology (Fixed/Unfixed). Dependent variable (y-axis): Mean support sum. A) Data size: 1000 characters, B) Data size: 2500 characters. Error bars indicate 95% confidence intervals.

Both methods return higher support values when the sample size is increased from 1000 to 2500 characters, but Bayesian posterior probabilities increase more, despite them being substantially higher also for 1000 characters. Maximum likelihood bootstrapping is more conservative and less powerful than Bayesian inference. Support for the nodes in the total evidence tree was lower with maximum likelihood bootstrapping and 2500 characters, than with Bayesian inference and only 1000 characters.

Bayesian inference clearly seems to be more sensitive to under-parameterization and as a result it probably returns too high support values in many analyses of empirical data.

Paper III

In paper III we analysed a matrix of 33,149 molecular characters to infer the cpDNA phylogeny of the tribe *Sileneae* (Caryophyllaceae), the largest amount of data ever used for inferring phylogenies on this taxonomic level of Angiosperms. The taxon sampling included in total twenty-four species and all recognized genera. We were able to confidently resolve many, previously poorly resolved, phylogenetic relationships. Twelve of the twenty-one internal branches in the phylogeny receive maximum support in the analysis of the total character matrix (Fig. 3) and these branches are also found in most of the partitioned data sets. The genera *Atocion*, *Viscaria*, *Eudianthe*, and *Heliosperma* form a strongly supported group positioned as sister to a large clade consisting of all sampled species of *Silene* and *Lychnis*, and thus leaving *Petrocoptis* and *Agrostemma* outside.

A group represented by two very similar species, *Silene aegyptiaca* and *S. atocioides*, is found to be sister to all other species of *Silene* and *Lychnis*, causing the latter genus to be nested within *Silene* under its current circumscription. It is presently unclear if this position is unique to the chloroplast phylogeny, and thus possibly a result of an ancient interspecific hybridization. Despite the apparent morphological similarities and the strongly supported sister relationship in the molecular phylogeny, *Silene aegyptiaca* and *S. atocioides* have surprisingly different sequences throughout the investigated regions of cpDNA. We believe that it is likely that the cpDNA substitutions rate for some reason has been elevated in this group.

The position of the strongly supported clade consisting of *Silene cryptoneura* and *S. sordida* is enigmatic. In most cpDNA partitions it is resolved as sister group to *Lychnis*, but in one of the partitions it is strongly supported as sister to *Silene* subgenus *Behen* (Paper III, Fig. 3). On the basis of nrDNA ITS data the two species do *not* group together, but are both nested *within* subgenus *Behen* (Oxelman & Lidén, 1995). The apparent hard incongruence between cpDNA and nuclear data, and possibly also the conflicting signals between the cpDNA partitions, suggest a complicated past of this group.

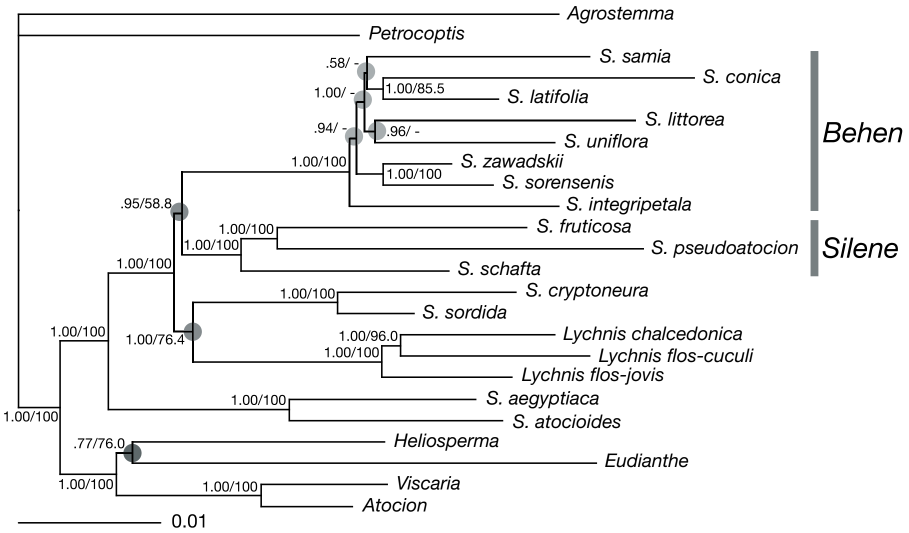

Figure 3. Total evidence phylogeny based on nucleotides and indel characters. Numbers on nodes are branch support (Bayesian posterior probabilities / parsimony bootstrap). Only values above 0.50 /50% are shown. Vertical bars indicate the two subgenera in *Silene*. Gray dots indicate nodes failing to converge to high support values discussed in the text.

The situation in *Silene* subgenus *Behen* is complicated. The monophyly of the group on the basis of cpDNA data is strongly supported in all the sequenced regions and a long branch separates it from the rest of *Sileneae*, but only two of the internal branches receive strong support with both Bayesian inference and parsimony. The estimated internal branch lengths are very short and the terminal branches indicate unequal substitution rates.

A common argument for using e.g. maximum likelihood as optimality criterion rather than parsimony is that the latter method is inconsistent under some conditions (e.g. long branch attraction, LBA). The consistency, ascribed to maximum likelihood, is a desirable statistical property and means that the method will converge on the correct solution with infinite amount of data. One problem, easy to forget, is that likelihood methods only are consistent under the correct model of evolution. In analyses of empirical data the correct model of sequence evolution is, of course, unknown, and the model used in the inference is by necessity a simplification of the actual substitution process. Despite large amount of data several nodes in the phylogeny, not only in *Behen*, could not be confidently resolved. Some of these relationships did not seem to benefit from the addition of more data (Paper III, Fig. 5).

Suzuki et al. (2002) showed that when the underlying topology is a star tree Bayesian inference can give high support to an arbitrary resolution. Lewis et al. (2005) argued that the cause of this effect is that the Bayesian prior distribution puts no weight on unresolved topologies and showed that short branches (near true polytomies) can be receive high support in standard MCMC implementations. They showed that by allowing unresolved topologies to have some prior probability and use a reversible-jump MCMC (not implemented in MrBayes) some short branches with previous high posterior probabilities (but low bootstrap frequencies) collapse. In our cpDNA phylogeny (paper III) some branches with low bootstrap support are even shorter than those highly sensitive to topology priors in Lewis et al. (2005), but because our study is based on more sequence data, the expected number of substitutions along branches with high posterior probabilities is generally larger.

Under the assumption of the chloroplast genome as a single evolving unit, the only possible explanation for incongruent results found in different partition of cpDNA is methodological artefacts, caused by, e.g. short internal branches and substitution rate heterogeneity. We explored the possibility of ancient interspecific chloroplast recombination as an alternative explanation to the apparent conflict among different parts of the sequenced data (Paper III, Fig. 3:4d and Fig. 7).

If the estimated short internal branches in *Behen* are the result of rapid radiation, we expected the high degree of homoplasy found in the group to be randomly distributed along the sequence. This does not seem to be the case and the dramatic effect on support values of excluding some taxa fits poorly with the idea of rapid radiation as the explanation for the observed pattern. The data are inconclusive, but it seems worthwhile to increase the taxon sampling from the subgenus *Behen* in order to more thoroughly examine the hypothesis of a reticulate past of its chloroplast genome.

In our investigations of chloroplast sequence evolution in *Sileneae* three taxa appear more enigmatic than the rest. *Lychnis chalcedonica* (front cover), *Silene conica*, and *Silene fruticosa*, representing three separate lineages, all exhibit extreme sequence evolution in two regions, physically separated by >10 kb of normal sequence, the *clpP1* gene region (Paper IV) and the *accD-psaI* gene region. These regions were excluded before the investigations referred to above were conducted. *Lychnis chalcedonica* has multiple, at least partial, copies in the *accD-psaI* region and the copy with putatively functional copies of the genes *accD* and *psaI* is relocated in the chloroplast genome. The two *Silene* species have a single copy of the region found in a position homologous to other species, but all three species have numerous large unique indels in the *accD* gene, none of which cause stop-codons.

In molecular plant systematics, investigators have strived to identify regions in the genome that have an appropriate level of variation for the questions asked. This is often done on the basis of other researchers experience and even on the availability of published primers for PCR and sequencing. Most studies are based on one or a few DNA regions in the size range of

500-1500 bp. To resolve relationships among closely related taxa and to better understand the evolutionary processes of sequence evolution there is a great need for more sequence data from non-model organisms. Some authors focus on comparisons of different short regions of non-coding cpDNA to find highly variable regions or even the "holy grail" of cpDNA variability (Shaw et al., 2005). Sure, there are differences in substitution rates in different regions, and some coding genes are, because of their highly constrained sequences, almost useless for lower level phylogenetic inference, but the rate differences in non-coding cpDNA is much more restricted and often lineage specific. A better strategy is to apply the long-range PCR approach used in our study of *Sileneae* (Paper III). Also when the sole purpose of sequencing is to yield as many potentially informative characters as possible to resolve a species phylogeny, sequencing of long continuous regions can reduce the effort, because the effect of primer mismatch are minimized and potential mix-ups or contaminations are more easily spotted. Our study also shows the potential for finding paralogous copies of cpDNA regions and to identify cases of genome rearrangements.

Paper IV

During the extensive sequencing for our study of chloroplast DNA evolution and phylogeny in *Sileneae* (Paper III) we discovered strikingly elevated substitution rates in the exon sequences of the *clpP1* gene in three *Sileneae* lineages (sect. *Conoimorpha*, *Lychnis*, and *Silene fruticosa*). The substitution rates of the magnitude found for some taxa in this study are the most extreme we are aware of from the chloroplast genome. For example, the uncorrected pairwise distance between closely related *Silene latifolia* and *S. conica* is 0.31, whereas the distances between *Silene latifolia* and *Amborella* or *Pinus* are 0.13 and 0.27, respectively.

Because the three lineages within *Sileneae* that display extreme substitution rates are only distantly related, the extreme increase in substitution rates probably has arisen multiple times. We surveyed all available sequences of *clpP1* on GenBank and found *Oenothera elata* to be suspiciously divergent. Not only did it show signs of elevated substitution rates, it also lacked introns in the gene, a feature shared with some of the *Sileneae* species with extreme substitution rates. We sequenced the *clpP1* gene region of four arbitrarily chosen species of *Oenothera* from different section in the genus and found highly elevated substitution rates in all of them. One of them, *Oenothera flava*, did have introns and was less divergent.

Because the exons in both *Sileneae* and *Oenothera* appeared functional (conserved reading frame and no stop-codons), despite high variability and several indels, we wanted to investigate if the gene was under selection. Branches both in *Sileneae* and *Oenothera* were found to have experienced highly significant positive selection with up to six times more non-synonymous than synonymous substitutions (Fig. 4).

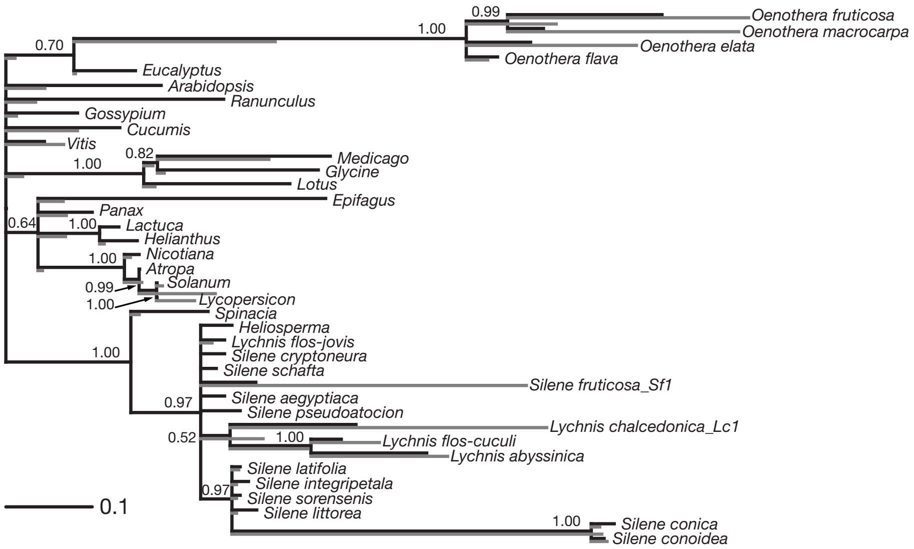

Figure 4. dS-tree based on the topology from Bayesian analysis of third positions from the *clpP1* exons. The dN-branch lengths are shown in gray. Numbers on nodes are Bayesian posterior probabilities (Bpp). Only Bpp values above 0.50 are shown.

Records of genes that exhibit positive selection as a whole and with non-synonymous substitutions more or less evenly spread are rare and, to our knowledge, there are no published cases of significant positive selection acting on chloroplast genes.

In the two species in *Sileneae* that display significant positive selection on their terminal branches (*Silene fruticosa* and *Lychnis chalcedonica*) we found multiple partial copies of the gene indicating several rounds of gene duplications possibly predating elevated substitution rates and positive selection. Previous studies have found that duplicated genes evolve at a faster rate (Otha, 1993), and that duplications can correlate with positive selection (Van de Peer et al., 2001).

The substitution rates of *clpP1* introns (e.g. in *Silene fruticosa*) do not seem to be different from any other non-coding cpDNA, and yet the exons evolve many times faster. This observation in combination with the estimated very high substitution rates for some taxa indicates that the *clpP1* gene has experienced increased mutation rate as a whole.

In the branch leading to *Silene conica* and *S. conoidea* the estimated dN/dS-ratio is close to one, indicative of neutral evolution. The absence of stop codons and frame-shifts, despite extreme variability and multiple indels, speak against lost functionality. Because only the cumulative pattern of nucleotide substitutions can be studied, an alternative explanation for the observed pattern can be that the lineage has experienced strong positive selection in the past, but the trace of this has been erased by more recent negative selection (Van de Peer et al., 2001). Denser taxon sampling could be a way to explore this possibility.

In *Lychnis,* a single intron loss event, ancestral to three extant species, is strongly supported, yet *L. chalcedonica* contain copies both with and without introns, indicating an ancient duplication. Other genes present in the duplicated fragment do not show signs of elevated substitution rates or positive selection. It is enigmatic that the apparently most recent duplication in *L. chalcedonica*, resulting in the complete copy without introns found as an insertion in another cpDNA region, show the most obvious signs of pseudogenization. The more ancient copies with introns appear to have changed considerably less, at least one of them. Further taxon sampling is, also in this case, a promising approach for better understanding of these exciting evolutionary processes. A preliminary hypothesis of the *clpP1* gene evolution in *Lychnis* is presented in figure 5.

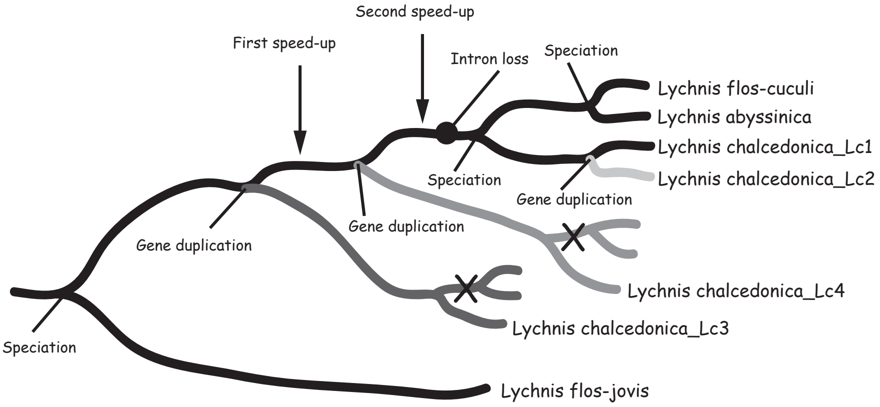

Figure 5. Preliminary hypothesis of the *clpP1* gene evolution in *Lychnis*. X indicate undetected or lost gene copy. The functional copy is found in the black lineage and always positioned between the *rps12* and *psbB* genes.

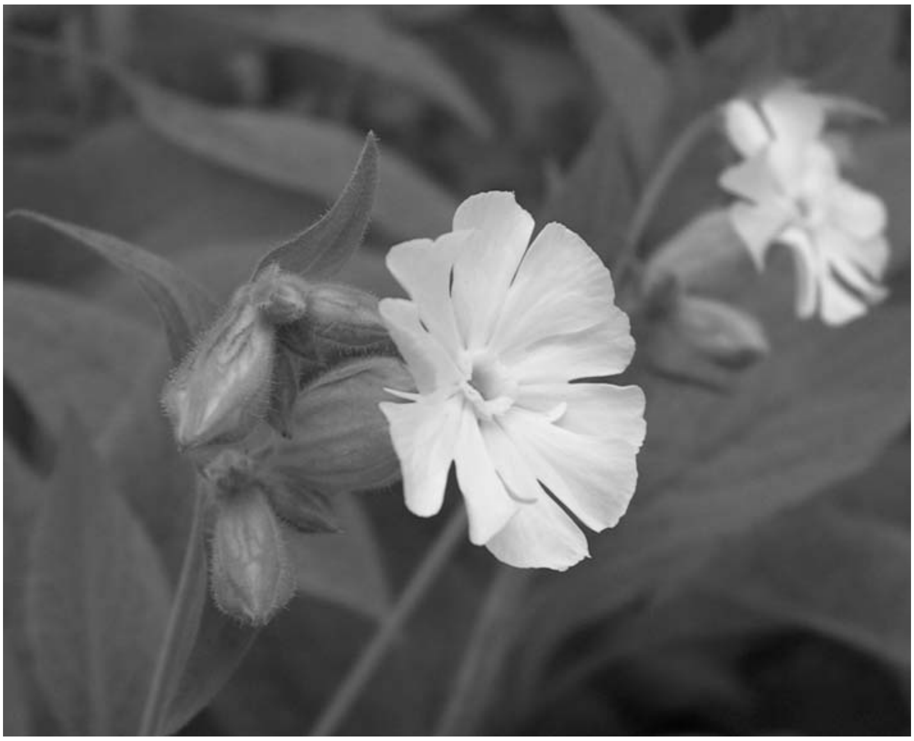

Figure 6. Silene latifolia Poir., Uppsala, Sweden.

Sammanfattning (Swedish summary)

Evolution handlar om förändring över lång tid. All förändring har sin grund i mutationer i det genetiska anlagen (organismers DNA). Somatiska mutationer, som t.ex. cancer, ärvs inte till kommande generationer, men de mutationer som sker i könscellerna återfinns i avkomman. Många förändringar har ingen betydelse för individens överlevnad, andra påverkar negativt och några få leder till bättre förutsättningar. Evolutionen har genom årmiljonerna gett upphov till ett släktträd, eller en fylogeni, som i teorin är helt dikotomt förgrenat, dvs varje gren delar förr eller senare upp sig i två nya.

Artbildning sker genom att populationer (från samma art) blir separerade från varandra så att inget utbyte av genetiskt material sker mellan dem. Denna process är i princip helt oberoende av förändringen av de genetiska anlagen genom mutationer.

En gren i det evolutionära släktträdet kan vara ouppdelad länge, eller så kan artbildningen ske väldigt snabbt (korta grenar). Förändringshastigheten i olika grenar kan skilja sig åt, vilket gör att två evolutionära linjer som delar en och samma anfader kan uppvisa olika grad av förändring trots att de med nödvändighet haft lika lång tid på sig att utvecklas. Många evolutionära grenar är utdöda idag. Fylogenin representerar en process som vi vill få kunskap om, men de data som vi har att tillgå är en ögonblicksbild av de arter som finns idag. Inom fylogenetisk rekonstruktion är det denna koppling mellan mönster och process man vill finna.

Kort kan man säga att min avhandling handlar om två centrala delar i fylogenetisk rekonstruktion, dels hur hypoteser genererade med olika analysmetoder skiljer sig åt och dels hur de data som vi huvudsakligen använder oss av för att estimera fylogenier i blomväxtsystematik är beskaffade.

Inom molekylärsystematik använder man sig i första hand av sekvensdata (ACGT) från organismernas DNA för att kartlägga hur de är släkt med varandra. Växter har förutom genomet i cellens kärna (kromosomerna) även två, betydligt mindre, cirkulära genom i cellernas mitokondrier respektive kloroplaster. Modern växtsystematik har huvudsakligen baserats på data från kloroplastgenomet. För ett drygt tjugotal blomväxter har hela kloroplastgenomet sekvenserats (ungefär 150,000 baspar). Ofta baseras dock släktskapsstudier på något eller några få tusen baspar. En fördel med sekvensdata från kloroplasten är att de antas sakna flertalet av de problem som vanligtvis är förknippade med det många tusen gånger större och betydligt sämre undersökta kärngenomet. Bland annat nedärvs kloroplasterna vanligen endast via moderslinjen och avsaknaden av rekombination gör att artbildning genom hybridisering (vanligt hos växter) inte ställer till problem i analyserna. Vanligtvis, men inte alltid, förekommer sekvenser från kloroplasten bara i en kopia, vilket förstås minimerar risken för att sekvensera "fel" kopia.

De två första artiklarna i avhandlingen handlar om hur olika mått på styrka för grenar i ett fylogenetiskt träd skiljer sig åt. Vissa grenar i släktträdet underbyggs av mycket data och är därför ganska tillförlitliga, medan andra relationer är mer osäkra och resultaten från olika analysmetoder kan då skilja sig åt.

Denna del av min forskning baseras dels på datorsimuleringar, men också på analys av de stora mängder sekvensdata som jag tagit fram för artikel **III**. Fördelen med simuleringar är att man använder sig av en evolutionär modell och ett givet utgångsträd för att generera sina data, vilket gör att man alltid vet vilket det sanna släktträdet är. Nackdelen är att simuleringar aldrig fullt ut kan modellera de komplexa processer som genererat de verkliga sekvenserna.

I artikel **I** undersöker jag hur två olika mått på styrka eller stöd, Bayesianska a posteriori-sannolikheter och Maximum Likelihood bootstrap-frekvenser, hos grenar i fylogenetiska träd skiljer sig åt. Detta gjordes genom att först simulera DNA-sekvenser utifrån en viss fylogeni under en viss evolutionär modell, för att sedan analysera dessa data med de två olika metoderna. Dels görs analyserna med samma modell som använts för att generera data, men också med en modell som är betydligt förenklad. Genom detta tillvägagångssätt kan de två metodernas estimeringsfel uppskattas, både när modellen är den rätta och när den är fel (underparameteriserad). En stor fördel med simuleringar är just att det sanna släktträdet är känt samt att slumpfel kan minimeras genom att simuleringarna upprepas ett stort antal gånger. Vi kan också kontrollera graden av underparameterisering. Slutsatserna av studien var att metoderna tydligt skiljer sig åt, den Bayesianska metoden ger generellt högre stödvärden än bootstrap-metoden för sanna grenar i trädet. Det är förstås en stor fördel att sanna grupper ges högre tilltro, men tyvärr var denna effekt kopplad till högre stödvärden även för falska grupper. Vid analys med korrekt modell var denna effekt liten, men vid underparameterisering blev den besvärande hög (grupper med uppskattade a posteriori-sannolikheter mellan 95 och 100 % var fel vid 16% av tillfällena). Vi fann också att när datamängden var stor blev skillnaden i stöd mellan grupperna liten, betydligt mindre än tidigare observationer gjorda på verkliga data.

Artikel **II** handlar just om skillnaderna mellan metoderna när verkliga data analyseras. Problemet med jämförelser baserat på verkliga data är att man vanligtvis inte har möjlighet till upprepning, dessutom är estimeringsfelet svårt att skatta eftersom det sanna släktträdet är okänt. Genom att utgå från en mycket stor mängd data (delvis samma som i artikel **III**) och beräkna det mest troliga släktträdet baserat på hela datamängden erhölls en referens mot vilken analyser av delar av datasetet kunde jämföras. Även i denna studie analyserades dataseten, både det fullständiga och de små partitionerna, med båda metoderna. Analyserna gjordes med både den enklaste och den mest komplicerade tillgängliga evolutionära modellen, men även den senare visade sig dåligt förklara de observerade datamönstren. Bootstrap-metoden var mycket robust mot modellvalet, men ineffektiv och konservativ, medan den Bayesianska metoden återigen gav generellt högre stödvärden, också för grenar som inte återfanns i referensträdet.

För artikel **III** sekvenserade jag ca 30,000 baspar från kloroplastgenomet för 24 arter i blomväxttribusen *Sileneae*. I *Sileneae* ingår bland annat vitblära (figur 6), rödblära, smällglim och backglim från släktet *Silene*, gökblomster och studentnejlika (omslagsbilden) från släktet *Lychnis*, tjärblomster från släktet *Viscaria*, klätt från släktet *Agrostemma* och alpglim (figur 7) från släktet *Heliosperma*. Gruppen omfattar totalt ungefär 700 arter i huvudsak förekommande på norra halvklotet (knappt tjugo arter i Sverige) och är en del av familjen nejlikväxter (Caryophyllaceae).

Analyserna gav mycket starkt stöd för flera, tidigare okända eller dåligt stödda, släktskapsrelationer. Ett av de mest spännande resultaten var att en liten grupp glimmar från östra medelhavet (representerade av *Silene aegyptiaca* och *S. atocioides*) hamnade som systergrupp till alla andra representanter från släktena *Silene* och *Lychnis*. De två arterna är morfologiskt väldigt lika, men visade sig vara förvånansvärt olika på sekvensnivå. Två andra arter från sydvästra Turkiet (*Silene cryptoneura* och *S. sordida*) bildar en starkt stödd grupp, men data från olika delar av kloroplastgenomet resulterar i två olika placeringen av dem i trädet. Detta är mycket förbryllande eftersom kloroplastgenomet normalt betraktas som en enda evolverande enhet som nedärvs på mödernet, vilket innebär att även om arter har uppstått genom hybridisering, alltså att artträdet inte är dikotomt förgrenat utan mer som ett nätverk, så ska kloroplastträdet förbli "normalt". Placeringarna av gruppen i trädet skiljer sig också tydligt från tidigare analyser baserat på sekvensdata från kärngenomet. Detta är en indikation på att gruppen kan ha ett hybridursprung.

En annan del av trädet som betedde sig gåtfullt var åtta arter från undersläktet *Behen* i *Silene*. Gruppen som helhet har mycket starkt stöd, men upplösningen inom gruppen är dålig trots den, i sammanhanget, stora mängden data. En förklaring till situationen är att artbildningen inom gruppen en gång i tiden troligen skedde under, evolutionärt sett, kort tid samt att förändringshastigheterna i de olika grenarna har varit olika sedan dess. Detta gör estimeringen av relationerna besvärlig. Det är dock mycket konflikt i data och den verkar inte vara slumpmässigt fördelad. En möjlig, men osannolik, förklaring till de mönster vi ser både i *Behen* och när det gäller positionen av *Silene cryptoneura* och *S. sordida* skulle kunna vara att kloroplaster överförts med pollen när två arter hybridiserat så att två olika kloroplasttyper hamnat i samma växt (heteroplasmi) och därefter rekombinerat (bytt anlag med varandra). Det förefaller inte som mer sekvensering av de arter som ingår i studien skulle kunna bringa klarhet i frågan, men om fler arter från gruppen analyserades skulle denna spektakulära hypotes bättre kunna utvärderas.

En fantastisk fördel med att jobba med verkliga data är att man ibland snubblar över resultat man aldrig ens kunnat drömma om. Tre arter som återfinns i helt olika delar av släktträdet, *Silene conica* (sandglim), *Silene fruticosa* (en medelhavsart som skulle kunna kallas buskglim på svenska) och *Lychnis chalcedonica* (studentnejlika), visade sig tidigt vara lite av problembarn på molekylärlabbet och jag förbannade, mer än en gång, att just de valts ut att ingå i studien. Särskilt två regioner av kloroplastgenomet (kring *accD*-genen och kring *clpP*-genen), åtskillda av mer än 10,000 baspar av mer oproblematisk sekvens, var särskilt konstiga hos de tre arterna. Hos studentnejlika saknades mer än två tusen baspar där *accD*-genen borde finnas och i dess ställe fanns nästa tre tusen baspar från en helt annan del av kloroplastgenomet. När denna del undersöktes visade sig att regionerna helt enkelt bytt plats, dessutom fann jag ytterligare tre delkopior av regionen (dock med okänd hemvist). I den enda potentiellt fungerande kopian av *accD*-genen hade det skett omfattande förändringar, stora delar hade tillkommit (ibland hundratals baspar) och andra saknades helt. Liknande mönster återfanns i sekvenserna från *Silene conica* och *S. fruticosa* indikerande att alla tre upplevt stor dynamik i *accD*-genen. Situationen för de tre arterna i den andra genen, *clpP,* var så extrem att det blev en egen artikel (**IV**).

Eftersom många gener bildar viktiga protein leder mutationer i dessa regioner oftare till negativa konsekvenser för individen och därmed förs sådana mutationer vidare i mindre utsträckning. Detta gör att sekvensskillnaden mellan olika arter normalt är betydligt mindre för gener jämfört med DNA som inte har någon uppenbar funktion. Sekvenserna från *clpP*-genen för "problembarnen" visade sig vara väldigt förändrade, t.ex. var skillnaden större mellen sandglim och vitblära (dess syster-art) än mellan tall och vitblära. Alla tre arterna uppvisade relativa förändringshastigheter aldrig tidigare skådat för DNA från kloroplastgenomet.

Vissa gener är uppdelade i flera proteinkodande delar (exoner) separerade av icke-kodande delar (introner). Nästan alla undersökta blomväxter, utom gräsen, har introner i *clpP*-genen.

Protein är uppbyggda av aminosyror. Det är DNA:t som bestämmer vilka aminosyror som ska ingå i proteinet genom att dess baser översätts tre och tre till aminosyror (t.ex. ger CAA aminosyran glutamin). Eftersom det finns 64 (4^3) möjliga tripletter (kodon), men bara 20 aminosyror kodas de flesta aminosyror av mer än ett kodon (även CAG ger glutamin). Förändringar i den genetiska koden som *inte* ändrar aminosyrasammansättningen i ett protein kallas synonyma substitutioner. För nästan alla gener är synonyma substitutioner många gånger vanligare än icke-synonyma substitutioner (jämför resonemanget ovan). I enstaka fall kan en förändring av en aminosyra (icke-synonym mutation) leda till bättre förutsättningar för individen. Ju större förbättringen är, desto troligare är det att den fixeras i populationen, dvs efter många generationer har alla individer den. Adaptiv eller positiv selektion innebär att de icke-synonyma substitutionerna är fler än de synonyma i en utvecklingslinje (gren i fylogenin). Det finns idag många exempel på enskilda kodon i DNA-sekvenser som uppvisar detta fenomen, men räknat över en hel gen är exemplen få. Positiv selektion från kloroplastgenomet har mig veterligen aldrig tidigare dokumenterats.

Tre av de extremt långa grenarna i genfylogenin (omslagsbilden) av *clpP*-genen uppvisar dock signifikant positiv selektion mätt över genen som helhet. Grenen som leder till buskglim (*Silene fruticosa*) har sex gånger fler icke-synonyma substitutioner. Hos studentnejlika (*Lychnis chalcedonica*), som också uppvisar positiv selektion, finns *clpP*-genen (eller delar av den) i fyra kopior, men bara en är funktionell. Två av kopiorna, som är resultatet av en relativt sentida duplikation, saknar introner i genen precis som artens två närmaste släktingar *L. flos-cuculi* och *L. abyssinica*. De två andra kopiorna i *L. chalcedonica* har introner. Analys av sekvenserna från *Lychnis*-arterna visar att intronförlusten skedde före artbildningen inom släktet och att kopiorna med introner är resterna av mycket gamla duplikationer.

Sannolikt hade jag inte upptäckt dessa märkligheter om jag valt att sekvensera många korta regioner av DNA, som andra forskare tidigare utnyttjat (den "normala" strategin), istället för långa sammanhängande fragment, som till övervägande del aldrig tidigare använts för fylogenetisk rekonstruktion. Molekylärsystematisk forskning är otroligt spännande!

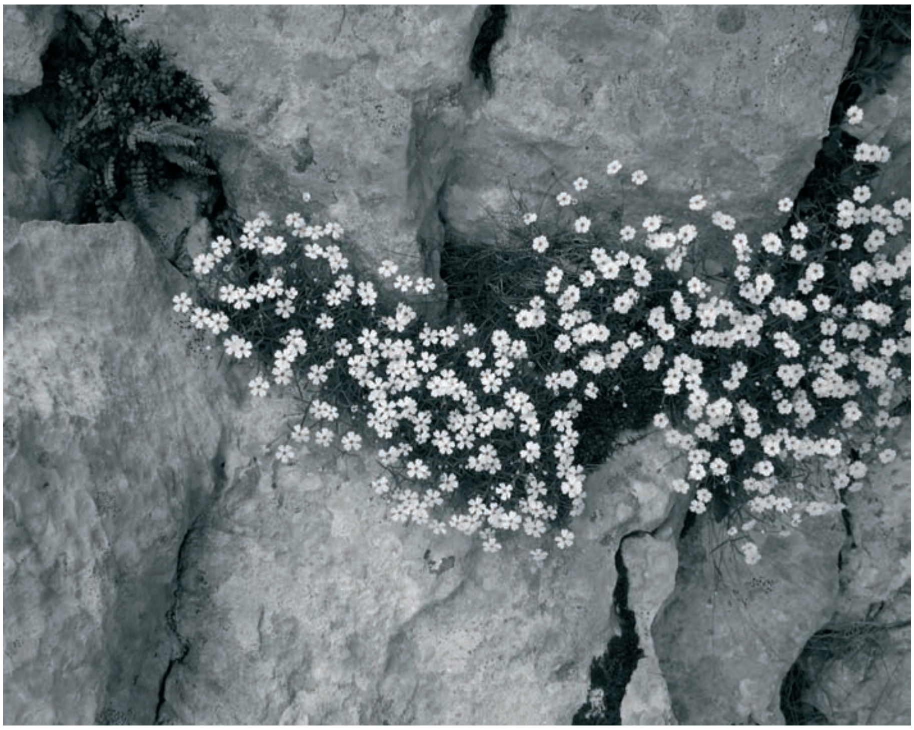

Figure 7. Heliosperma pusillum Vis., Kamniske Alps, Slovenia.

Tack! (Acknowledgements)

Så var det till slut dags att skriva tacket till alla som gjort denna avhandling möjlig. Risken att jag glömmer någon är tyvärr överhängande, men jag ska göra mitt bästa.

Bengt, vilken fantastisk handledare du har varit för mig. Du har många egenskaper jag verkligen värderar högt: integritet, ödmjukhet, engagemang, fokus och tillgänglighet. Diskussionerna med dig kräver att man skärper sig, men jag upplever alltid att du lyssnar. Oklara tankar och frön till ideér gror med uppmuntran. Du gav mig kanske oförsiktigt fria tyglar på labbet, ett tag var spåren väldigt många, men det ordnade sig ju till slut och många spännande saker finns kvar att göra. Utan din snabba och klockrena feedback hade inte heller denna avhandling varken blivit klar i tid eller särskilt bra. Min förhoppning är att du även fortsättningsvis har lust att vara min mentor, för än har jag mycket kvar att lära av dig.

Ett stort tack vill jag rikta till Birgitta och Kåre Bremer, Leif Tibell, Mats Thulin, Katarina Andreasen, Niklas Wikström samt Inga och Olle Hedberg. Jag vill också tacka Tom Britton och Bodil Svennblad för ett lärorikt och givande samarbete och Mikael Thollesson för support i alla former.

Att jag fastnade för växtsystematik från början berodde nog till stor del på den härliga stämningen på "alg & moss" som jag upplevde som student. Tänk om jag fick bli doktorand där! Det blev ju lättare sagt än gjort eftersom jag var klar med examensarbetet precis när institutionen antagit fyra nya doktorander. Det tyckte jag var missflyt då, men tänk vad trist doktorandtiden blivit utan er som kollegor. Särskilt vill jag tacka Annika för allt skoj i och utanför vårt rum, tänk vad tiden går fort när man har roligt, och Poppen, för att du är en helschysst gubbe med skön humor.

Alla nuvarande kollegor förtjänar ett varmt tack: Anja, Anneleen, Bozo, Cajsa, Catarina, Frida, Hugo, Il-Chan, Rikke, Sanja och Sunniva. Plus en drös med mer eller mindre gamla botaniska kollegor: Jesper, Kornhall, Dick, Hobbe, Maria, Björn-Axel, Elisabeth, Johanne, Kristina, Starri, Anders, Christina, Jenny, Torsten, Helena, Ghebre, Lars-Gunnar, Mats, Ulf, Bengt, Anders,...

Many thanks to Doug Stone and Sylvain Razafimandimbison for stimulating discussions.

Det finns också många trevliga zoologer som jag vill tacka: Johan (för alla fylogenetiska insikter), Hege, Karolina, Andreas (för klustret), Mattias, Per, Isabel, Fredrik och alla andra nuvarande eller föredettingar på systzoo.

Nahid vill jag särskilt tacka, för allt du lärt mig på labb, för alla trevliga pratstunder och all uppmuntran. Tack också Ulla, Agneta och Afsaneh.

Mina föräldrar förtjänar ett stort tack för både arv och miljö. Min bror, mina svärföräldrar, mina svågrar och svägerskor, tack ni också.

Tack alla andra vänner som förgyllt tillvaron, och under senaste sommaren gjort livet lite roligare för en hårt prövad familj med en pappa som hellre jobbade än semestrade.

Mina underbara barn, Hugo, Sara och Ivar, vad vore livet utan er? Det har känts ganska hårt att jag nu i sluttampen inte kunnat ägna er all den tid ni förtjänar, eller som Sara sa: "min pappa jobbar alla dagar". När jag någon gång känt mig lite nere har ni funnits där för att med all kraft påpeka att livet innehåller annat än avhandlingsarbete. Jag instämmer, särskilt nu.

Slutligen vill jag tacka dig Elsa, för att du är den finaste och mest fantastiska människa som finns - jag älskar dig!

References

Andolfatto, P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437: 1149-1152.

APG. 1998. An ordinal classification for the families of flowering plants. Annals of the Missouri Botanical Garden 85: 531-553.

APG II. 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Botanical Journal of the Linnean Society 141: 399–436.

Brandley, M.C., Schmitz, A., and Reeder, T.W. 2005. Partitioned Bayesian Analyses, Partition Choice, and the Phylogenetic Relationships of Scincid Lizards. Systematic Biology 54: 373-390.

Brandley, M.C., Leache, A.D., Warren, D.L., and McGuire, J.A. 2006. Are unequal clade priors problematic for Bayesian phylogenetics? Systematic Biology 55: 138-146.

Bennett, M.D., and Leitch, I.J. 1997. Nuclear DNA Amounts in Angiosperms - 583 New Estimates. Annals of Botany 80: 169-196.

Bremer, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42: 795–803.

Bremer, K. 1994. Branch support and tree stability. Cladistics 10: 295–304.

Buckley, T.R. 2002. Model misspecification and probabilistic tests of topology: Evidence from empirical data sets. Systematic Biology 51:509-523.

Cummings, M. P., Handley, S. A., Myers, D. S., Reed, D. L., Rokas, A., and Winka, K. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Systematic Biology 52: 477–487.

Douady, C.J., Delsuc, F., Boucher, Y., Doolittle, W.F., and Douzery, E.J.P. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Molecular Biology and Evolution 20: 248–254.

Efron, B., Halloran, E., and Holmes, S. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA 93: 7085–7090.

Farris, J.S., Albert, V.A., Källersjö, M., Lipscomb, D., and Kluge, A.G. 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12:99–124.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology 27: 401-410.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Associates, Inc., Massachusetts.

Goremykin, V.V., Hirsch-Ernst, K.I., Wölfl, S., and Hellwig, F.H. 2003. Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. Molecular Biology and Evolution 20: 1499-1505.

Graur, D., and Li, W.-H. 2000. Fundamentals of Molecular Evolution, 2nd ed. Sinauer Associates, Inc., Massachusetts.

Hillis, D.M., and Bull, J.J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biology 42: 182–192.

Holder, M., and Lewis, P.O. 2003. Phylogeny estimation: Traditional and Bayesian approaches. Nature Reviews Genetics 4: 275-284.

Huelsenbeck, J.P., Rannala, B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Systematic Biology 53: 904-913.

Huelsenbeck, J.P., and Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17: 754-755.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R., and Bollback, J.P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294: 2310–2314.

Karol, K.G., McCourt, R.M., Cimino, M.T., and Delwiche, C.F. 2001. The closest living relatives of land plants. Science 294: 2351-2353.

Larget, B., and Simon, D.L. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular Biology Evolution 16: 750-759.

Leaché, A.D., and Reeder, T.W. 2002. Molecular systematics of the eastern fence lizard *Sceloporus undulatus*: A comparison of parsimony, likelihood, and Bayesian approaches. Systematic Biology 51: 44-68.

Lemmon, A.R., and Moriarty, E.C. 2004. The Importance of Proper Model Assumption in Bayesian Phylogenetics. Systematic Biology 53: 265–277.

Lewis, P.O., Holder, M.T., and Holsinger, K.E. 2005. Polytomies and Bayesian Phylogenetic Inference. Systematic Biology 54: 241–253.

Li, S. 1996. Phylogenetic Reconstruction Using Markov Chain Monte Carlo. Ph.D. Thesis. Ohio State University.

Mau, B. 1996. Bayesian Phylogenetic Inference via Markov Chain Monte Carlo Methods. Ph.D. Thesis, University of Wisconsin.

Murphy, W.J., E. Eizirik, S.J. O'brien, O. Madsen, M. Scally, C.J. Douady, E. Teeling, O.A. Ryder, M.J. Stanhope, W.W. De Jong, and M.S. Springer. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294: 2348–2351.

Müller, K.F., Borsch, T., and Hilu, K.W. 2006. Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: Contrasting matK, trnT-F, and rbcL in basal angiosperms. Molecular Phylogenetics and Evolution 41: 99–117.

Nylander, J.A.A., F. Ronquist, J. P. Huelsenbeck, and Nieves Aldrey, J.-L. 2004. Bayesian Phylogenetic Analysis of Combined Data. Systematic Biology 53: 47-67.

Ohta, T. 1993. Pattern of Nucleotide Substitutions in Growth Hormone-Prolactin Gene Family: A Paradigm for Evolution by Gene Duplication. Genetics 134: 1271 - 1276.

Oxelman, B., and Lidén, M. 1995. Generic Boundaries in the Tribe *Sileneae* (Caryophyllaceae) as Inferred from Nuclear rDNA Sequences. TAXON 44: 525-542.

Pickett, K.M., and Randle, C.P. 2005. Strange bayes indeed: uniform topological priors imply non-uniform clade priors. Molecular Phylogenetics and Evolution 34: 203–211.

Randle, C.P., and Pickett, K.M. 2006. Are Nonuniform Clade Priors Important in Bayesian Phylogenetic Analysis? A Response to Brandley et al. Systematic Biology 55: 147-151.

Randle, C.P., Mort, M.E., and Crawford, D.J. 2005. Bayesian inference of phylogenetics revisited: developments and concerns. TAXON 54: 9-15.

Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. Journal of Molecular Evolution 43: 304–311.

Shalchian-Tabrizi, K., Minge, M.A., Cavalier-Smith, T., Nedreklepp, J.M., Klaveness, D., and Jakobsen, K.S. 2006. Combined Heat Shock Protein 90 and Ribosomal RNA Sequence Phylogeny Supports Multiple Replacements of Dinoflagellate Plastids. Journal of Eukaryotic Microbiology 53: 217–224.

Shaw, J., Lickey, E.B., Beck, J.T., Farmer, S.B., Liu, W., Miller, J., Siripun, K.C., Winder, C.T., Schilling, E.E., and Small, R.L. 2005. The Tortoise and the Hare II: Relative Utility of 21 Noncoding Chloroplast DNA Sequences for Phylogenetic Analysis. American Journal of Botany 92: 142–166.

Simmons, M.P., Pickett, K.M., and Miya, M. 2004. How meaningful are Bayesian support values? Molecular Biology and Evolution 21: 188-199.

Smedmark, J.E.E, Swenson, U., and Anderberg, A.A. 2006. Accounting for variation of substitution rates through time in Bayesian phylogeny reconstruction of Sapotoideae (Sapotaceae). Molecular Phylogenetics and Evolution 39: 706-721.

Strugnell, J., Norman, M., Jackson, J., Drummond, A.J., and Cooper, A. 2005. Molecular phylogeny of coleoid cephalopods (Mollusca : Cephalopoda) using a multigene approach; the effect of data partitioning on resolving phylogenies in a Bayesian framework. Molecular Phylogenetics and Evolution 37: 426-441.

Suzuki, Y., G. V. Glazko, and M. Nei. 2002. Overcredibility of molecular phylogenetics obtained by Bayesian phylogenetics. PNAS 99: 16138-16143.

Svennblad, B., Erixon, P., Oxelman, B., and Britton, T. 2006. Fundamental Differences Between the Methods of Maximum Likelihood and Maximum Posterior Probability in Phylogenetics. Systematic Biology 55: 116-121.

Van de Peer, Y., Taylor, J.S., Braasch, I., and Meyer, A. 2001. The Ghost of Selection Past: Rates of Evolution and Functional Divergence of Anciently Duplicated Genes. Journal of Molecular Evolution 53: 436-446.

Whittingham, L. A., B. Slikas, D. W. Winkler, and F. H. Sheldon. 2002. Phylogeny of the tree swallow genus Tachycineta (Aves: Hirundinidae), by Bayesian analysis of mitochondrial DNA sequences. Molecular Phylogenetics and Evolution 22: 430–441.

Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. Molecular Biology and Evolution 14: 717–724.

Yukawa, M., Tsudzuki, T., and Sugiura, M. 2006. The chloroplast genome of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*: complete sequencing confirms that the *Nicotiana sylvestris* progenitor is the maternal genome donor of *Nicotiana tabacum*. Molecular Genetics and Genomics 275: 367-373.

Zwickl, D.J., and Holder, M.T. 2004. Model Parameterization, Prior Distributions, and the General Time-Reversible Model in Bayesian Phylogenetics. Systematic Biology 53: 877–888.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 226

Editor: The Dean of the Faculty of Science and Technology

Distribution: publications.uu.se
urn:nbn:se:uu:diva-7166

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2006