Reconstructing Evolutionary History: Algorithmic, Statistical, and Computational Challenges

A comprehensive exam for the department of Computer and Information Science

Victor Hanson-Smith
with advising from Joe Thornton and John Conery

September 13, 2010

1 Introduction

Over the last four billion years, life emerged on Earth and evolved into a menagerie of diverse forms and functions (Fenchel, 2002; Knoll, 2004; Popa, 2004; Schopf, 2006). Evolutionary biologists investigate the underlying processes—mutation, drift, and natural selection—that give rise to this diversity (Ridley, 2004; Barton et al., 2007). It can be challenging to directly observe these processes, especially for biological systems that evolved over millions (or billions) of years. Biologists have traditionally used "stones and bones" (e.g. fossils) to infer evolutionary history, but the fossil record is incomplete and not easily searchable. It would be a frustrating state of affairs if all advancement in evolutionary biology relied on lucky shovels! Fortunately, gene sequencing technology provides alternatives to fossil-based inference; we can gather genetic sequences from contemporary species and then computationally infer their shared history. In the last decade, whole-genome sequencing has provided a nearly limitless supply of genetic data.

The field of computational phylogenetics is concerned with algorithms and models for inferring evolutionary history with confidence boundaries. Phylogenetic models make several simplifying assumptions about underlying evolutionary processes, and a significant body of work addresses these assumptions. In general, all proposed solutions to increase biological realism come at the expense of increased computational complexity. Contemporary evolutionary models have become sufficiently complex that software implementations require search heuristics and clever algorithm design. This tension between scientific realism and computational tractability is the central crisis of computational phylogenetics.

coloration. The dark moths were well-adapted to the sooty forest and were more likely to survive predation. Over time, evolution selected for the dark phenotype; the light phenotype declined into minority.

Phenotypes are encoded in genetic material (i.e. DNA), and emerge when genetic systems interact with their environment. Within each living cell, so-called "coding regions" of DNA are transcribed into RNA and then translated into proteins. Individual proteins and networks of proteins are expressed in patterns which determine an individual's phenotype; protein expression patterns are often tailored to an organism's environment in response to diet, stress, light periodicity, and temperature.

A species population occasionally becomes separated due to spatio-temporal barriers, including geographic separation and niche specialization. These now-divided subpopulations are then free to accumulate unique mutations such that the two subpopulations become genetically incompatible, and thus new species. A history of evolutionary lineage-splitting can be expressed as a type of tree graph, called a cladogram. The terminal nodes of a cladogram correspond to observed taxa; the internal branching pattern expresses the shared ancestry of the terminal nodes. Some cladograms are rooted, in which case the root node corresponds to the most-recent-common-shared ancestor of all taxa on the tree. Figure 1 illustrates a simple rooted cladogram for the history of the family *Hominidae*.

A phylogram is a special type of cladogram in which the branch lengths correspond to a measure of evolutionary distance, where distance is typically measured as the number of differences between molecular sequences. The primary goal of phylogenetic inference is to determine the correct phylogram for a given set of extant taxa.

ample begins with an ancestral sequence $NEDP$, which stands for the amino acids asparagine (N), glutamic acid (E), aspartic acid (D), and proline (P). Some evolutionary event—such as population isolation, or perhaps a genome duplication—separates $NEDP$ into two divergent lineages. In one lineage, N evolves to D and then E, giving rise to the intermediate ancestral sequence $EEDP$. In the other lineage, the character D evolves to V, giving rise to the intermediate ancestor $NEVP$. Subsequent evolutionary events cause both intermediate ancestors to undergo further duplication. In one descendant of $EEDP$, the second character E is deleted (as indicated by the state '—'). In the other three lineages, different mutations occur: P to threonine (T), P to alanine (A), and E twice mutates to glutamine (Q). This collection of mutation and duplication events yields four extant descendant sequences: EDP, $EEDT$, $NQVP$, and $NQVA$. Figure 2 explicitly shows the history as just described, but typically this lineage would be hidden from us if we collected the descendant sequences from the wild.

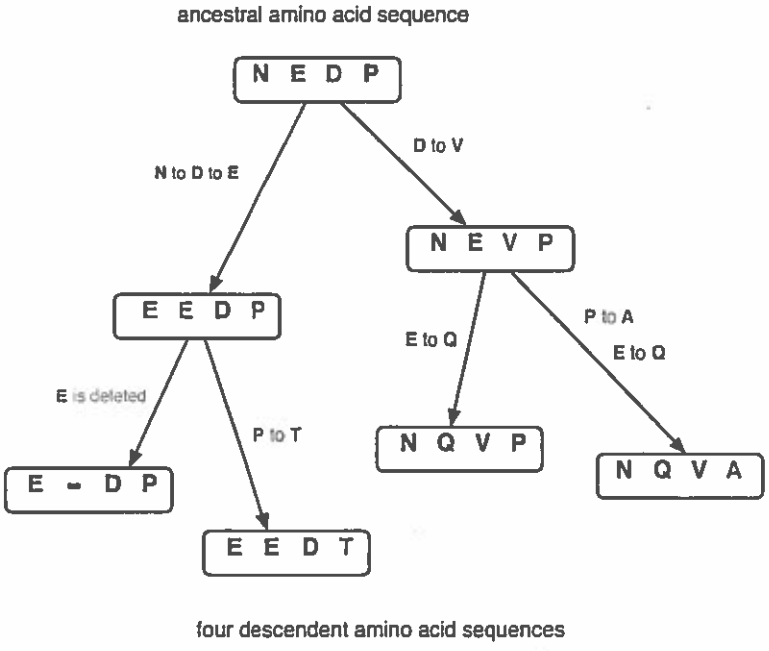

Figure 2: An ancestral amino acid sequence $NEDP$ evolves into four descendant sequences.

of DNA are inserted and deleted. Consequently, a particular gene might be encoded with x number of amino acids in one species, but $x + \Delta$ or $x - \Delta$ number of amino acids in another species. The placement of indels can drastically affect the alignment outcome. To demonstrate this point, consider several incorrect alignments for the sequences shown in Figure 2. First, consider an alignment in which all the characters are inferred to be non-homologous:

```
EDP----------
---EEDT--------
-------NQVP----
----------NQVA
```

Next, consider an alignment in which we correctly infer homology for characters one and four, but fail to fully align characters two and three:

```
E-D----P
EED----T
N----QVP
N----QVA
```

Finally, consider an alignment in which we incorrectly place the indel in character one:

```
-EDP
EEDT
NQVP
NQVA
```

In addition to these three examples, you can imagine many other alignments, each with a unique indel placement solution. An obvious question is *which alignment is the best?* A large body of bioinformatics research investigates algorithms for finding the best alignment. In the remainder of this section, I describe two basic alignment algorithms upon which most contemporary alignment methods are based.

include right, down, and diagonally down-right to other corners on F. All paths terminate in the lower-right corner. Figure 3 illustrates four pairwise alignments of sequences EDP to $EEDT$, and the corresponding paths across F. It should be noted that there exist many more alignment solutions than the four shown in Figure 3 . Here we consider the cost of these four alignment individually.

from the paths shown in Figure 3.

- alignment 1: $-5+5+6+1=7$
- alignment 2: $5-5+6+1=7$
- alignment 3: $5-5+6-5-5=-4$
- alignment 4: $-5\times 7=-35$

You can see the Needleman-Wunsch algorithm returns a score of 7 for both the first and second alignments. Indeed, we get the same score for placing an indel at either site 1 or 2 within the sequence EDP. The problem of indel placement, in this example, is problematic and unresolvable.

2.2 The Smith-Waterman Algorithm

Alignments generated using the Needleman-Wunsch algorithm can be poor quality for sequences with regions of great dissimilarity. For example, many protein sequences encode a multi-domain protein, where each domain has a different structure and biological function. The sequence sites between domains are often functionally unimportant and their sequences tend to evolutionarily drift. In many cases it is more biologically appropriate to locally align the highly conserved domains of protein sequences, rather than globally align all sites.

The Smith-Waterman algorithm extends the Needleman-Wunsch algorithm to locally align two sequences (Smith and Waterman, 1981). The Smith-Waterman algorithm changes all the negative values in the F matrix to 0. Consequently, alignments are not penalized for matching dissimilar residues, but are still rewarded for matching similar residues.

- COFFEE and its descendant T-COFFEE use a consistency-based function to incorporate the results from CLUSTAL and L-ALIGN into a single result (Notredame et al., 1998; Cedric Notredame, 2000).

- 3DCOFFEE extends T-COFFEE to incorporate three-dimensional protein structure information. 3DCOFFEE uses a technique called "threading" to map homologous sequences onto a single protein structure (O'Sullivan et al., 2004).

- Espresso is an extension of 3DCOFFEE, where the structural templates for sequence threading are automatically identified and retrieved from a database of protein structures (Armougom et al., 2006).

- MUSCLE uses the unweighted-pair-method-with-arithmetic-mean to build a progressive guide tree. MUSCLE refines the guide tree based on the results of the progressive alignement, and then repeats the alignment procedure. (Edgar, 2004)

- ProbCons implements progressive global alignment and uses a method called "probabilistic consistency" to update the score for matching residues x and y based on the triangulated score of matching x to some third residue z, and y to z (Do et al., 2005).

- DIALIGN provides an alternative to the Needleman-Wunsch and Waterman-Smith algorithms; DIALIGN compares regions of sequences rather than individual resides (Subramanian et al., 2008).

- PRANK implements global sequence alignment, and attempts to disambiguate between insertions and deletions by incorporating phylogenetic information (Loytynoja and Goldman, 2005, 2008).

where D_{JC} is the corrected distance and D_{obs} is the observed distance (the Hamming distance) between two sequences. The use of the term $\left(\frac{3}{4}\right)$ appears as a consequence of Jukes and Cantor assuming that all nucleotide substitutions can occur with equal probability. Kimura extended this idea to incorporate different substitution rates between different states. Specifically, Kimura's correction method incorporates a different rate for transitions ($A \leftrightarrow G$ and $C \leftrightarrow T$) and transversions ($A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$, $T \leftrightarrow G$) (Kimura, 1980).

When using distances—or some other metric—to cluster sequences into a tree, the order of the clustering can strongly affect the phylogenetic outcome. To deal with this complication, several phylogenetic algorithms use the distance-based tree as a starting point and then heuristically search for the best tree according to some optimality criterion. A simple criterion is to select the tree with the shortest sum of branch lengths because this tree implies the "minimum evolution" of characters on the tree (Saitou and Nei, 1987). Another criterion—the "least squares" method—seeks to minimize the squared difference between a tree's branch lengths and the pairwise distances in the distance matrix (Fitch and Margoliash, 1967).

A third criterion—called "maximum parsimony" (MP)—scores trees based on the number of mutations implied by their branching pattern; the tree implying the fewest mutations is chosen because it provides the simplest hypothesis for how the sequence data arose (Edwards and Cavalli-Sforza, 1963; Camin and Sokal, 1965; Kluge and Farris, 1969; Farris, 1970; Fitch, 1971; Farris, 1977). Maximum parsimony reached widespread popularity among evolutionary biologists, but it was shown to be statistically inconsistent under some conditions (Felsenstein, 1978). A method of phylogenetic inference is said to be consistent if the method converges on the true phylogeny as the length of the observed sequences increases. Inconsistent methods can pull us towards an incorrect phylogeny. Although MP yields the correct tree in many cases, Felsenstein showed that MP can

where μ is our assumed rate of evolution. The probability that any number of mutations—from zero to infinity—occur in time t can be calculated by summing over all values k:

$$P(0 \leq k \leq \infty | t) = \sum_{k=0}^{\infty} \frac{\mu t^k e^{-t\mu}}{k!} = 1.0 \tag{3}$$

Expression 3 is used to calculate the likelihood of a single phylogenetic branch for a single sequence site as follows.

Suppose we observe an evolutionary character—a single nucleotide or an amino acid—currently in some state x, where x is one of the letters in the nucleotide or amino acid alphabet. Also suppose we have a matrix R expressing the relative substitution rates between states. R is an n-by-n matrix, where n is the size of the alphabet. Finally, we have a vector π expressing the expected frequencies of each state. Putting all these elements together, x will mutate to state y over time t with probability calculated as follows:

$$P(x \rightarrow y | t) = \sum_{k=0}^{\infty} (\pi_x \pi_y R_{xy}^k) \frac{t^k e^t}{k!} \tag{4}$$

. . . where R_{xy} is the relative rate of x transitioning to y, and (R_{xy}^k) the extrapolated rate of $x \rightarrow y$ occurring over k steps. π_x and π_y are the frequencies of states x and y, otherwise known as the stationary frequencies. Expression 4 is typically shown in a more compact form:

$$P(t) = \sum_{k=0}^{\infty} \frac{Q\mu t^k}{k!} = e^{Q\mu t} \tag{5}$$

. . . where the matrix Q equals $\Pi R - I$. Π is the diagonal matrix, where $\Pi[a, a]$ equals the equilibrium frequency π_a for state a in our alphabet. I is the identity matrix. The value μ is

likelihoods are calculated by deeper recursion to the branches descending from nodes u_1 and u_2. Eventually, the recurrence arrives at a leaf node u_T. The partial likelihood $L_y^{u_T}$ of state y at node u_T equals 1.0 if u_T is state y in the sequence data; otherwise $L_y^{u_T}$ equals 0. Figure 4 illustrates the data structures involved in this recursion.

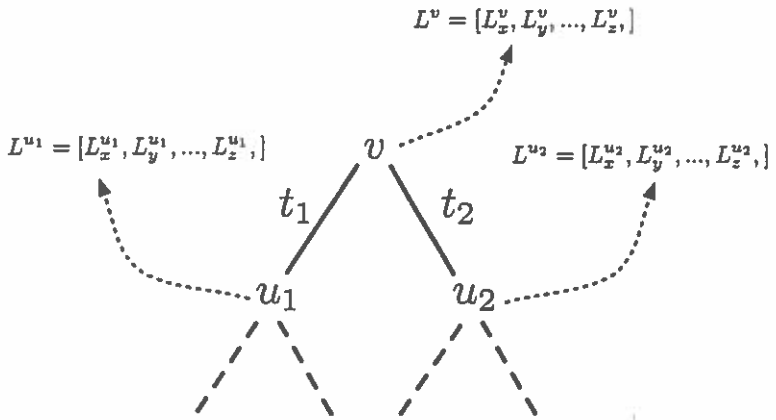

Figure 4: The recursive data structure for the likelihood algorithm. We pick an arbitrary root node v, with descendant branches t_1 and t_2 leading to nodes u_1 and u_2. We recursively calculate a vector of partial likelihoods for each node on the tree. For example, the vector L^{u_1} contains the partial likelihood $L_x^{u_1}$ of node u_1 existing as state x, the partial likelihood $L_y^{u_1}$ of u_1 existing as state y, etc.

3.2 Likelihood Methods Make Critical Simplifying Assumptions.

The basic phylogenetic Markov model makes three simplifying assumptions. First, the model assumes site independence: every site in the sequence alignment is assumed to evolve independently of other sites. This assumption allows for simplicity in calculating likelihoods (Expression 6), but this assumption is known to be unrealistic for protein-coding sequences. The structural integrity of a protein relies on covalent and electrostatic bonds between amino acids; these bonds often occur between amino acids that are geographic neighbors in a three-dimensional protein structure, but are far apart in a one-dimensional sequence. The assumption of site independence is inaccurate in this case because some sites are evolutionarily constrained by their interactions in a three-dimensional

the probability of sampling state y from the stationary distribution and transitioning to state x. In other words: $\pi_x P(x \rightarrow y|t) = \pi_y P(y \rightarrow x|t)$. Time reversibility greatly simplifies the likelihood calculation because the algorithm becomes independent of the tree's root position; time reversibility allows us to choose any arbitrary root when applying the likelihood recursion (Expression 8). A proof for this simplification is found in Felsenstein's so-called "pulley principle" (Felsenstein, 1981).

The assumption of time reversibility has been shown to be biologically inaccurate in mitochondrial DNA because the heavy DNA strand spends more time in the single-strand state (in which there is higher probability of mutation) during DNA replication (Faith and Pollock, 2003). A similar violation occurs in nuclear DNA during replication, in which the nontemplate DNA strand experiences an excess of $C \rightarrow T$ and $G \rightarrow A$ mutations (Polak and Arndt, 2008). In order to eliminate the assumption of time reversibility, an extended Markov model has been proposed (Barry and Hartigan, 1987) and implemented in Java (Jayaswal et al., 2005). However, this implementation calculates likelihoods only on single given tree and does not provide facilities to search the space of possible topologies. Indeed, ML optimization is more computationally-intensive without time reversibility because Felsenstein's pulley principle is no longer valid and multiple rootings must be considered on every tree.

3.3 Nucleotide Models

The likelihood method requires a matrix of relative substitution rates between states: this is the matrix R introduced in Expression 4. Nucleotide substitution rates are usually estimated as free parameters from the sequence data. Several nucleotide matrices have been proposed:

- Jukes and Cantor's model, JC69, assumes equal mutation probabilities between all nucleotides (Jukes and Cantor, 1969).

- Jones, Taylor, and Thornton calculated the JTT matrix using the same approach used for Dayhoff matrices, but with a much larger database of proteins (Jones et al., 1991).

- Adachi and Hasegawa calculated the mtREV matrix specifically for vertebrate mitochondrial proteins. The authors used a maximum likelihood approach—rather than a counting approach—to find the substitution matrix that optimized the likelihood of a phylogeny relating all the available vertebrate mitochondrial proteins (Adachi and Hasegawa, 1996).

- Whelan and Goldman used an ML method (as previously pioneered for the mtREV matrix) to calculate the WAG matrix for 3095 globular protein sequences from 182 protein families (Whelan and Goldman, 2001).

- Le and Gascuel improved Whelan and Goldman's method by incorporating variable evolutionary rates in the phylogeny of their curated proteins. Le and Gascuel used their method to estimate the LG matrix from approximately 50000 protein sequences (Le and Gascuel, 2008).

3.5 Heterogeneity

The basic likelihood algorithm, as described in Section 3.1.2, assumes the evolutionary process is homogenous across all sequence sites and lineages. This assumption is probably incorrect for most datasets: it has been widely observed that evolutionary rates vary across sites (Fitch and Margoliash, 1967; Uzzell and Corbin, 1971), and across lineages (Lopez et al., 2002; Philippe et al., 2003). Several models have been proposed to incorporate various forms of evolutionary heterogeneity (Fitch and Markowitz, 1970; Galtier, 2001; Yang, 1994, 1996; Tuffley and Steel, 1997), but a particularly popular approach is to partition a sequence alignment into subalignments, and then apply a unique model to each partition (Yang, 1996; Ronquist and Huelsenbeck, 2003). This parti-

$$L(t, b, \pi, Q_1, Q_2, ..., Q_K, w|D) = \prod_i \sum_j P(D_i|t, b, \pi, Q_j)w_j \quad (10)$$

Another mixture model, the CAT model, incorporates heterogeneity about the stationary distribution by fitting a mixture of equilibrium state frequencies to data (Lartillot and Philippe, 2004). The CAT model provides a mixture of π-vectors. Each sequence site i is fit to one of the π-vectors. An additional parameter c stores the π assignments, such that π_{c_i} is the π-vector to which site i is assigned. The CAT model calculates the likelihood $L(t, b, Q, \pi_1, \pi_2, ..., \pi_K, c|D)$ of a tree t, with branch lengths b, substitution process Q, and a mixture of K π-vectors (Expression 11).

$$L(t, b, Q, \pi_1, \pi_2, ..., \pi_K, c|D) = \prod_i P(D_i|t, b, Q, \pi_{c_i}) \quad (11)$$

Strictly speaking, the CAT model is a *dynamically partitioned* model—not a mixture model—because each site is assigned to one of the π-vector mixtures rather than calculating the likelihood of each mixture for all sites.

A third mixture model—the branch length mixture model—incorporates a form of heterogeneity called "heterotachy" in which sites evolve according to lineage-specific rates (Kolaczkowski and Thornton, 2008; Meade and Pagel, 2008). A heterotachous model incorporates multiple branch length sets, where each set contains a unique length for every branch in the tree. The branch length mixture model calculates the likelihood $L(t, Q, \pi, b_1, b_2, ..., b_K, w|D)$ of tree t, rates Q, state frequencies π, and K branch lengths sets $b_1, b_2, ..., b_k$ (Expression 12).

$$L(t, Q, \pi, b_1, b_2, ..., b_K, w|D) = \prod_i \sum_j P(D_i|t, Q, \pi, \theta, b_j)w_j \quad (12)$$

methods to optimize continuous parameters sequentially. This approach, which I refer to as successive line maximization (SLM), optimizes a collection of parameters θ by assuming the likelihood function is "partially separable" such that the function $L(\theta_1, \theta_2, ..., \theta_N|D)$ can be optimized as smaller functions $L(\theta_1|D)$, $L(\theta_2|D)$, etc.

My own work suggests that SLM is inappropriate for ML inference because the assumption of partial separability is incorrect. If two or more parameters are dependent then they are not partial separable, and SLM is not guaranteed to return ML values. My initial observations suggest that mixture proportions ($[w_1, w_2, ..., w_K]$ in Expression 9) are extremely dependent with branch lengths and sGHC seems to be especially inappropriate for optimizing mixture models.

PAUP and PhyML implement a non-SLM technique, called the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method. BFGS is a quasi-Newton method that can optimize multiple dimensions simultaneously (Gill et al., 1982), assuming the multidimensional function can be approximated as a quadratic. Unfortunately, the BFGS implementations in PAUP and PhyML have serious limitations. PAUP is commercial software and the source code is not freely available; PAUP's implementation cannot be directly analyzed. PhyML's implementation of BFGS—available as open source code—can fail due to memory limitations; in these cases, PhyML recognizes the fault and switches to Brent's method. My own investigation of BFGS in PhyML suggests that BFGS reaches memory limits in almost all practical cases. This shortcoming of BFGS has been addressed in an algorithm called L-BFGS, which optimizes BFGS to use limited memory (Nocedal, 1980); surprisingly, this work has not yet been adopted by the phylogenetics community.

Quasi-Newton methods (including SLM and BFGS) assume the underlying function is unimodal. When this assumption is true, any uphill move is guaranteed to lead closer to the maximum likelihood value. If the function is multimodal, however, then quasi-Newton methods can become

$$T_{i+1} = T_i \times \frac{e^{\left[log(\frac{T_f}{T_0})\right]}}{\phi - 1.0} \tag{14}$$

. . . where T_i+1 is the next temperature, T_i is the current temperature, T_f is the final temperature (typically 0.0), T_0 is the initial temperature, and ϕ is the number of desired STA iterations.

STA (and stochastic approaches in general) are not widely used in a phylogenetic context to optimize continuous parameters. The first phylogenetic application of STA was to optimize continuous parameters for a heterotachous mixture model (Kolaczkowski and Thornton, 2008). My software (in development) extends this implementation, and to the best of my knowledge, is the only phylogenetic software that stochastically optimizes continuous parameters. My preliminary analysis suggests that STA finds higher optima than quasi-Newton methods, at the cost of increased computation time. STA is computationally expensive because it must calculate the likelihood of every proposal. In a situation where sGHC reaches an ML conclusion after several hundred propositions, STA can typically require millions of propositions. If it takes one second to compute the likelihood of a phylogeny, then sGHC would finish in under one hour, while STA would finish in 11 days.

3.6.2 Optimizing the Topology

Topologies are discrete structures defined by their branching pattern. The space of topologies for a given alignment is explored by swapping branches. The algorithm "nearest-neighbor interchange" (NNI) swaps branches that are adjacent (Moore et al., 1973), while the algorithm "subtree pruning and regrafting" (SPR) swaps branches from across the tree (Swofford, 2003). A discrete graph of topologies, related by single swaps, forms a nested set of Petersen graphs (Holton and Sheehan,

The goal of model selection is to find the evolutionary model that best-fits sequence data without *over*-fitting the data with too many parameters. For nucleotide data, the likelihood ratio test is typically used because nucleotide models are nested (Posada and Crandall, 1998). For amino acid data, the Akaike Information Criterion is typically used to select among non-nested models (Akaike, 1973).

Two models are "nested" if both contain the same terms and one of the models has at least one additional term. For example, the nucleotide model JC69 is nested inside the model K80. Whereas K80 includes separate terms α and β for transitions and transversions, JC69 assumes all rates are equal—essentially setting $\alpha = \beta$. The likelihood ratio test (LRT) compares the fit of a complex model to a nested model. The LRT calculates a test statistic $S = -2 \times ln\left(\frac{L_1}{L_2}\right)$, where L_1 is the maximum likelihood of the simpler model and L_2 is the maximum likelihood of the complex model. The probability of the test statistic can be approximated by consulting a chi-square distribution with degrees of freedom equal to the number of parameter terms not shared between the complex and simple models. If the resulting probability is statistically significant (typically measured as $p \leq 0.05$), then the complex model is indeed a better fit. The LRT for maximum likelihood nucleotide models is implemented in the software ModelTest (Posada and Crandall, 1998).

The Akaike Information Criterion (AIC) calculates a single score that can be used to select among non-nested models, which includes amino acid models (Akaike, 1973). The AIC score is calculated as $AIC = 2k - 2ln(L)$, where k is the number parameters in the model and L is the maximized likelihood of the model. When comparing AIC scores from two models, the lower score indicates a better fit. The AIC can intuitively be understood as penalizing the model for using multiple parameters, and rewarding the model for yielding a high likelihood score. AIC is implemented to test among amino acid models in the software package ProtTest (Abascal et al.,

times-prior is lower, the proposal is accepted with probability determined by the likelihood ratio $\left(\frac{P(\theta_2)}{P(\theta_1)} \times \frac{L(t,\theta_2|d)}{L(t,\theta_1|d)}\right)$ between the proposal and the current state, where $P(\theta_1)$ and $P(\theta_2)$ are the prior probabilities of the current state θ_1 and the proposed state θ_2. Using this approach, the Markov chain will spend a few iterations exploring low-likelihood parameter values, but will spend many iterations exploring high-likelihood values. The posterior probability of a particular parameter value can intuitively be understood as the amount of time the chain (or chains) spend dwelling on that value.

Bayesian methods are controversial because they require a prior probability distribution, and it is often unclear how the prior should be specified. Prior distributions for continuous parameters can be virtually any shape: exponential, binomial, logarithmic, etc. Flat priors—uniform over some distribution—are often used to reflect a lack of prior belief about a parameter. However, a uniform prior implies that we believe all parameter values are equally-likely; this is not the same as saying we know nothing about the parameter. The choice of prior has been shown to affect the accuracy of Bayesian methods: Kolaczkowski and Thornton observed that an exponential prior on branch lengths produced less accurate results than a uniform prior (Kolaczkowski and Thornton, 2007, 2009).

Regardless of priors, the Bayesian approach has been shown to be less accurate than ML in some cases. For example, the maximum *a posteriori* topology can be incorrect when the true evolutionary pattern contains particular signals of rate heterogeneity (Kolaczkowski and Thornton, 2007). Moreover, Bayesian branch lengths underestimate short branches; ML infers more branch lengths more accurately (Schwartz and Mueller, 2010).

Ultimately, ML and Bayesian methods fundamentally disagree about how to find the "best" phylogeny. The course of evolution occurred in exactly one way. ML seeks this single evolutionary

habitation—evolved independently on at least eighteen different occasions (Yokoyama et al., 2008). For more examples of historical ASR experiments, see (Thornton, 2004) and (Liberles, 2007).

In the early days of ASR, maximum-parsimony (MP) was the method *du jour* for reconstructing ancestral states: paleobiologists assigned states to ancestral nodes so as to minimize the number of state changes along the branches of the tree (Fitch, 1971; Hartigan, 1973) (See also (Swofford and Maddison, 1987, 1992; Maddison and Maddison, 1992)). MP was used to reconstruct ancestral lysozymes (Malcolm et al., 1990), the mouse L1 protein (Adey et al., 1994), the bovid ribonuclease (Stackhouse et al., 1990), and the artiodactyl ribonuclease (Jermann et al., 1995).

In the context of ASR, MP poses several problems. First, MP can yield several equally-best ancestral states at a given site, but provides no method for choosing the single-best state. This is troublesome if we are interested in chemically synthesizing ancestral molecules: the cost of manufacturing and investigating all the equally-best ancestral combinations can be prohibitively expensive. Second, when there exists asymmetry in transformation probabilities among states, MP can be systematically biased against changes from ancestral rare states to common extant states (Collins et al., 1994). Third, MP can produce biased reconstructions when the rate of evolution is not constant across the phylogeny (Cunningham et al., 1998). Finally, MP methods for ASR fail to incorporate information about branch lengths, mutation rates, or substitution rates. This means that a mutation from some state x to another state y is equally likely over branch lengths that are very short and very long.

4.1 An Empirical Bayesian Approach

As an alternative to MP, Yang et al. proposed an empirical Bayesian (EB) ASR method (Yang et al., 1995). Yang's method is Bayesian because it calculates posterior probabilities for ancestral

struction calculates the posterior probability of a state at a single ancestral node, and integrates over all possible ancestral states at all other nodes. A joint reconstruction calculates the posterior probability of states simultaneously for all ancestral nodes. Koshi and Goldstein introduced a dynamic algorithm for marginal ancestral reconstruction (Koshi and Goldstein, 1996), and Pupko et al. introduced a dynamic algorithm for joint reconstruction (Pupko et al., 2000). These algorithms perform with equivalent computational complexity, scaling linearly with the number of taxa. Yang implemented both variants in the software package PAML (Yang, 1997, 2007).

Marginal and joint reconstructions can yield disagreeing ancestral reconstructions. The appropriate method depends on the specific phylogenetic question being asked. For example, if we want to resurrect the maximum *a posteriori* ancestral sequence for only one ancestor, then a marginal reconstruction is appropriate. On the other hand, if we want to reconstruct the maximum *a posteriori* mutational trajectory—chains of mutations traveling through several ancestral nodes—then a joint reconstruction is appropriate.

4.2 Other Approaches

The empirical Bayesian (EB) approach (as described in section 4.1) assumes the alignment, tree, model, and model parameters are known *a priori* to be correct. In practice, this assumption is often not valid; for many real-world datasets, alternatives to the ML tree and parameter values cannot be ruled out. Bayesian methods have been proposed to accommodate these sources of uncertainty. Pagel et al. proposed a Bayesian method for integrating topological uncertainty into inference of ancestral states for binary and other discrete characters (Pagel et al. 2004). Schultz and Churchill proposed a Bayesian method to integrate uncertainty about the parameters of the evolutionary model into discrete character reconstructions (Schultz and Churchill 1999). For inference of an-

which an alignment algorithm returned an unbelievable result. Alignment algorithms are known to struggle when pairwise identities are low (Rost, 1999), but little else is known about how often and in what conditions alignment algorithms fail because there has been no systematic accounting for alignment successes and failures across diverse molecular sequence domains. This problem is compounded by the fact that virtually all alignment algorithms are scored for accuracy against the benchmark BaliBase (Thompson et al., 2005), but otherwise there exists a paucity of standardized alignment benchmarks. Alignment algorithms have evolved to be good at solving Balibase, but what about other datasets?

Aside from sequence alignments, the accuracy of phylogenetic Markov models is limited by their simplifying assumptions (as I described in section 3.2). For example, the assumption of equilibrium homogeneity is violated when one or more species experience unique environmental constraints on their nucleotide or amino acid composition. This situation can arise when a population colonizes a colder environment, thus relaxing selective pressures for a high equilibrium proportions of nucleotides guanine and cytosine (because G-C bonds are more thermostable than A-T bonds). As another example, the assumption of substitution process homogeneity is violated by the fact that some protein sites have specific amino acid states that are critical for structural integrity—and thus these sites experience strong evolutionary conservation—while other sites experience relaxed constraints on their state. Both of these simplifying assumptions—equilibrium homogeneity and substitution process homogeneity—can be overcome with linage-specific and site-specific mixture models (as discussed in section 3.5).

However, other violations are more difficult to overcome. The assumption of site independence is violated by any pair of protein sequence sites that interact when a protein folds into its three-dimensional tertiary structure. For example, two threonine resides in the penicillopepsin protein

processor unit (Charalambous et al., 2005), the IBM BlueGene/L (Ott et al., 2007), and the IBM Cell processor (Stamatakis et al., 2007). In all cases they achieved significant speedup, but these improvements remain off-limits for evolutionary biologists unwilling to purchasing exotic hardware like GPUs and Cell processors; the maintenance and operation of non-traditional hardware is often beyond the capabilities of most molecular biology labs. That said, many labs do support clusters with mainstream architectures (such as multicore chips from Intel and AMD). Consequently, I suspect that breakthroughs in high-performance phylogenetics will continue to come from further development of multiprocessor algorithms that use existing libraries—MPI and OpenMP—to operate in generic multiprocessor environments (Suchard and Rambaut, 2009).

5.3 Proposed Research

In light of the computational phylogenetics problems I outlined over the last 40+ pages, I propose three research goals.

5.3.1 Finding the best optimization algorithm

Over the last year, I implemented a branch length mixture model (as described in section 3.5) within PhyML's C code. I observed that PhyML's successive line maximization (SLM) method (as described in section 3.6.1) struggled to optimize the parameters of our mixture model, especially when I used more than two mixture components. I subsequently implemented a simulated annealing heuristic as an alternative to successive line maximization methods. My software is called "PHYESTA" (pronounced "fiesta") and stands for **Phy**ML extended for simulated thermal annealing. The source code for PHYESTA can be found online here:
[http://markov.uoregon.edu/software/phyesta].

is the likelihood surface multimodal or unimodal? A small body of work has shown that the likelihood surface can be multimodal, even with simple models and small datasets (Steel, 1994; Chor et al., 2000). In light of this work, it remains unclear if multimodality is the common case for most phylogenetic datasets, or if multimodality is the minority conditiion.

5.3.2 Reducing the algorithmic complexity of the likelihood function

The second project I propose is to reduce the complexity of the recursive algorithm for calculating likelihoods (as described in expression 4 of secion 3.1.2). Current implementations of the likelihood algorithm spend the majority of CPU clocktime calculating and retrieving partial likelihoods during a post-order traversal of the phylogeny. The complexity of the traversal can be reduced by using the tree structure to identify redundant calculations. Current software removes totally-redundant columns as a pre-processing step, and I propose that we can use the tree to remove partially-redundant columns. For example, consider this alignment:

```
seq1 ATTGTGA
seq2 TAAGTGA
seq3 CTGTGCA
seq4 CGGGCCT
```

Consider the tree ((seq1,seq2),(seq3,seq4)). On the ancestral node (seq1,seq2), the subcolumn "TA" occurs twice, and the subcolumn "GG" occurs twice. The likelihood algorithm can be optimized by memo-izing the first partial likelihood for "TA" and reusing it later. This optimization, which I call "tree-based subalignment compression," could significantly speedup likelihood calculations at the cost of a small increase in memory consumption.

diverse selective constraints.

6 Conclusion

Phylogenetic inference is a necessary precursor to understand evolutionary history, which is essential for ecological and biomedical progress (Barton et al., 2007). Early phylogenetic approaches made many simplifying assumptions about biological processes, but new insights into molecular biology—combined with computational advances—allow for models of molecular evolution to incorporate unprecedented scientific realism. Contemporary evolutionary models are becoming sufficiently complex that software implementations require search heuristics and clever algorithm design. This tension between scientific realism and computational tractability is the central crisis of computational phylogenetics; collaborations between biologists and computer scientists are necessary now more than ever before.

Charalambous, M., Trancoso, P., and Stamatakis, A. (2005). Initial experiences porting a bioinformatics application to a graphics processor. In *Advances in Informatics*, volume 2746 of *Lecture Notes in Computer Science*, pages 0302–9743. Springer Berlin / Heidelberg.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. (2003). Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Research*, 31(13):3497–3500.

Chor, B., Hendy, M. D., Holland, B. R., and Penny, D. (2000). Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Molecular Biology and Evolution*, 17(10):1529–41.

Collins, T. M., Wimberger, P. H., and Naylor, G. J. P. (1994). Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Systematic Biology*, 43(4):482–496.

Cunningham, C. W., Omland, K. E., and Oakley, T. H. (1998). Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology and Evolution*, 13(9):361–366.

Das, A. and Charkrabarti, B. K., editors (2005). *Quantum Annealing and Related Optimization Methods*. Lecture Notes in Physics. Springer.

Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5:345–352.

Do, C. B., Mahabhashyam, M. S., Brudno, M., and Batzoglou, S. (2005). Probcons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15(2):330–340.

Eck, R. and Dayhoff, M. (1968). *Atlas of protein sequence and structure*. National Biomedical Research Foundation.

Edgar, R. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

Edwards, A. and Cavalli-Sforza, L. (1963). The reconstruction of evolution. *Annals of Human Genetics*, 27:105–106.

Faith, J. J. and Pollock, D. D. (2003). Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics*, 165(2):735–45.

Farris, J. S. (1970). Methods for computing Wagner trees. *Systematic Zoology*, 19(1):83–92.

Farris, J. S. (1977). Phylogenetic analysis under dollo's law. *Systematic Zoology*, 26:77–88.

Felsenstein, J. (1978). Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology*, 227:401–410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

Fenchel, T. (2002). *The Origin and Early Evolution of Life*. Oxford University Press, Illustrated edition.

Huang, X. and Miller, W. (1991). A time-efficient linear-space local similarity algorithm. *Advances in Applied Mathematics*, 12(3):337–357.

Huelsenbeck, J. P. and Bollback, J. P. (2001). Empirical and heirarchical bayesian estimation of ancestral states. *Systematic Biology*, 50(3):351–366.

Huelsenbeck, J. P. and Ronquist, F. (2001). Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):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*, 294:2310–3214.

Jayaswal, V., Jermiin, L. S., and Robinson, J. (2005). Estimation of phylogeny using a general markov model. *Evolutionary Bioinformatics Online*, 1:62–80.

Jermann, T. M., Opitz, J. G., Stackhouse, J., and Benner, S. A. (1995). Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature*, 374:57–59.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1991). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, 8(3):275–282.

Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. In Munro, M., editor, *Mammalian Protein Metabolism, Volume III*, pages 21–132. Academic Press, New York.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120.

Kirkpatrick, S., Jr., C. G., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Kluge, A. G. and Farris, J. S. (1969). Quantitative phyletics and the evolution of anurans. *Systematic Zoology*, 18(1):1–32.

Knoll, A. H. (2004). *Life on a Young Planet.* Princeton University Press, 2004, illustrated edition.

Kolaczkowski, B. and Thornton, J. W. (2007). Effects of branch length uncertainty on bayesian posterior probabilities for phylogenetic hypotheses. *Molecular Biology and Evolution*, 24(9):2108–2118.

Kolaczkowski, B. and Thornton, J. W. (2008). A mixed branch length model of heterotachy improves phylogenetic accuracy. *Molecular Biology and Evolution*, 25(6):1054–1066.

Kolaczkowski, B. and Thornton, J. W. (2009). Long-branch attraction bias and inconsistency in bayesian phylogenetics. *PLoS ONE*, 4(12):e7891.

Koshi, J. M. and Goldstein, R. A. (1996). Probabilistic reconstruction of ancestral protein sequences. *Journal Molecular Evolution*, 42:313–320.

Lanave, C., Preparata, G., Sacone, C., and Serio, G. (1984). A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20(1):86–93.

Nocedal, J. (1980). Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782.

Notredame, C., Holm, L., and Higgins, D. G. (1998). COFFEE: an objective function for multiple sequence alignments. *Bioinformatics*, 14(5):407–422.

Ogden, T. H. and Rosenberg, M. S. (2006). "multiple sequence alignment accuracy and phylogenetic inference". *Systematic Biology*, 55(2):314–328.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., and Notredame, C. (2004). 3dcoffee: Combining protein sequences and structures within multiple sequence alignments. *Journal of Molecular Biology*, 340(2):385–395.

Ott, M., Zola, J., Aluru, S., and Stamatakis, A. (2007). Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. In *Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis*.

Pagel, M. and Meade, A. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology*, 53(4):571–581.

Pagel, M., Meade, A., and Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology*, 53(5):673–684.

Pauling, L. and Zuckerkandl, E. (1963). Chemical paleogenetics: molecular "restoration studies" of extinct forms of life. *Acta Chemica Scandinavia*, A(17):S9–S16.

Philippe, H., Casane, D., Gribaldo, S., Lopez, P., and Meunier, J. (2003). Heterotachy and functional shift in protein evolution. *IUBMB Life*, 55(4-5):257–65.

Polak, P. and Arndt, P. F. (2008). Transcription induces strand-specific mutations at the 5' end of human genes. *Genome Res*, 18(8):1216–23.

Popa, R. (2004). *Between necessity and probability: searching for the definition and origin of life.* Springer, Berlin.

Posada, D. and Crandall, K. A. (1998). Modeltest: testing the model of dna substitution. *Bioinformatics*, 14(9):817–818.

Pupko, T., Shamir, I. P. R., and Graur, D. (2000). A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular Biology and Evolution*, 17(6):890–896.

Ridley, M. (2004). *Evolution.* Blackwell Publishing.

Rodrigue, N., Lartillot, N., Bryant, D., and Philippe, H. (2005). Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene*, 347(2):207–17.

Rodrigue, N., Philippe, H., and Lartillot, N. (2006). Assessing site-interdependent phylogenetic models of sequence evolution. *Molecular Biology and Evolution*, 23(9):1762–75.

Rogers, J. S. (1997). On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Systematic Biology*, 46(2):345–357.

Swofford, D. L. (2003). Phylogenetic analysis using parsimony (and other methods) 4.0 b10. Sineer Associates, Inc.

Swofford, D. L. and Maddison, W. P. (1987). Reconstructing ancestral character states under Wagner parsimony. *Mathematical Biosciences*, 87:199–229.

Swofford, D. L. and Maddison, W. P. (1992). Parsimony, character-state reconstructions and evolutionary inferences. In R.L.Mayden, editor, *Systematics, Historical Ecology, and North American Freshwater fishes*, chapter 5. Stanford University Press.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680.

Thompson, J. D., Koehl, P., Ripp, R., and Poch, O. (2005). Balibase 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, 61(1):127–36.

Thornton, J. W. (2004). Resurrecting ancient genes: Experimental analysis of extinct molecules. *Nature*, 5:366–375.

Tuffley, C. and Steel, M. (1997). Modeling the covarion hypothesis of nucleotide substitution. *Math Biosci*, 147(1):63–91.

Uzzell, T. and Corbin, K. W. (1971). Fitting discrete probability distributions to evolutionary events. *Science*, 172(988):1089–96.

Wang, L. and Jiang, T. (1994). On the complexity of multiple sequence alignment. *J Comput Biol*, 1(4):337–48.

Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, 18(5):691–9.

Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science*, 319:416–417.

Worth, C. L., Gong, S., and Blundell, T. L. (2009). Structural and functional constraints in the evolution of protein families. *Nature Reviews Molecular Cell Biology*, 10(10):709–720.

Yang (1996). Maximum-likelihood models for combined analyses of multiple sequence data. *J Mol Evol*, 42(5):587–96.

Yang, Z. (1994). Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Systematic Biology*, 43(3):329–342.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, 13(5):555–556.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.