An Introduction to Reconstructing Ancestral Genomes

Lior Pachter

Abstract. Recent advances in high-throughput genomics technologies have resulted in the sequencing of large numbers of (near) complete genomes. These genome sequences are being mined for important functional elements, such as genes. They are also being compared and contrasted in order to identify other functional sequences, such as those involved in the regulation of genes. In cases where DNA sequences from different organisms can be determined to have originated from a common ancestor, it is natural to try to infer the ancestral sequences. The reconstruction of ancestral genomes can lead to insights about genome evolution, and the origins and diversity of function. There are a number of interesting foundational questions associated with reconstructing ancestral genomes: Which statistical models for evolution should be used for making inferences about ancestral sequences? How should extant genomes be compared in order to facilitate ancestral reconstruction? Which portions of ancestral genomes can be reconstructed reliably, and what are the limits of ancestral reconstruction? We discuss recent progress on some of these questions, offer some of our own opinions, and highlight interesting mathematics, statistics, and computer science problems.

1. What is comparative genomics?

These notes summarize a lecture at a special session of the American Mathematical Society on mathematical biology, during which we discussed the central problem of comparative genomics, namely how to reconstruct the ancestral genomes that evolved into the present-day extant genomes. This is fundamentally a statistics problem, because with a few exceptions, it is not possible to sequence the genomes of ancestral species, and one can only infer ancestral genomes from the multitude of genomes that can be sampled at the present time. The problem is a grand scientific challenge that has only begun to be tackled in recent years, now that whole genomes are being sequenced for the first time. Our aim is to introduce the reader to the statistical (and related mathematical) elements of the methods of comparative genomics, while providing a glimpse of the exciting results that are emerging from first generation attempts to reconstruct ancestral genomes. Due to the complex interdisciplinary scope of the subject, we have been forced to omit a lot of detail and many interesting topics, but we hope that the curious mathematical reader may find some threads worthy of further exploration.

We begin with a concrete example of one sequence from a single genome: the 16S ribosomal RNA (rRNA) gene from Salmonella typhimurium LT2. This sequence can be downloaded from the NCBI website at `http://www.ncbi.nlm.nih.gov/` by searching for the accession number "AE008857". The sequence is

```
   1 aattgaagag tttgatcatg gctcagattg aacgctggcg gcaggcctaa cacatgcaag
  61 tcgaacggta acaggaagca gcttgctgct tcgctgacga gtggcggacg ggtgagtaat
 121 gtctgggaaa ctgcctgatg gaggggggata actactggaa acggtggcta ataccgcata
 181 acgtcgcaag accaaagagg gggaccttcg ggcctcttgc catcagatgt gcccagatgg
 241 gattagcttg ttggtgaggt aacggctcac caaggcgacg atccctagct ggtctgagag
 301 gatgaccagc cacactggaa ctgagacacg gtccagactc ctacgggagg cagcagtggg
 361 gaatattgca caatgggcgc aagcctgatg cagccatgcc gcgtgtatga agaaggcctt
 421 cgggttgtaa agtactttca gcggggagga aggtgttgtg gttaataacc gcagcaattg
 481 acgttacccg cagaagaagc accggctaac tccgtgccag cagccgcggt aatacggagg
 541 gtgcaagcgt taatcggaat tactgggcgt aaagcgcacg caggcggtct gtcaagtcgg
 601 atgtgaaatc cccgggctca acctgggaac tgcattcgaa actggcaggc ttgagtcttg
 661 tagagggggg tagaattcca ggtgtagcgg tgaaatgcgt agagatctgg aggaataccg
 721 gtggcgaagg cggccccctg gacaaagact gacgctcagg tgcgaaagcg tggggagcaa
 781 acaggattag ataccctggt agtccacgcc gtaaacgatg tctacttgga ggttgtgccc
 841 ttgaggcgtg gcttccggag ctaacgcgtt aagtagaccg cctggggagt acggccgcaa
 901 ggttaaaact caaatgaatt gacggggggcc cgcacaagcg gtggagcatg tggtttaatt
 961 cgatgcaacg cgaagaacct tacctggtct tgacatccac agaactttcc agagatggat
1021 tggtgccttc gggaactgtg agacaggtgc tgcatggctg tcgtcagctc gtgttgtgaa
1081 atgttgggtt aagtcccgca acgagcgcaa cccttatcct ttgttgccag cgattaggtc
1141 gggaactcaa aggagactgc cagtgataaa ctggaggaag gtggggatga cgtcaagtca
1201 tcatggccct tacgaccagg gctacacacg tgctacaatg gcgcatacaa agagaagcga
1261 cctcgcgaga gcaagcggac ctcataaagt gcgtcgtagt ccggattgga gtctgcaact
1321 cgactccatg aagtcggaat cgctagtaat cgtggatcag aatgccacgg tgaatacgtt
1381 cccgggcctt gtacacaccg cccgtcacac catgggagtg ggttgcaaaa gaagtaggta
1441 gcttaacctt cgggagggcg cttaccactt tgtgattcat gactggggtg aagtcgtaac
1501 aaggtaaccg taggggaacc tgcggttgga tcacctcctt acct
```

This gene is a key component of the protein synthesis machinery in cells. In fact, it is such a basic ingredient of life, that the same sequence, albeit with minor modifications, exists in the genomes of all living organisms. For example, the reader is encouraged to compare it to the 16S rRNA gene in Escherichia coli (K12):

```
   1 aaattgaaga gtttgatcat ggctcagatt gaacgctggc ggcaggccta acacatgcaa
  61 gtcgaacggt aacaggaaga gcttgcttc tttgctgacg agtggcggac gggtgagtaa
 121 tgtctgggaa actgcctgat ggagggggat aactactgga aacggtagct aataccgcat
 181 aacgtcgcaa gaccaaagag ggggaccttc gggcctcttg ccatcggatg tgcccagatg
 241 ggattagcta gtaggtgggg taacggctca cctaggcgac gatccctagc tggtctgaga
 301 ggatgaccag ccacactgga actgagacac ggtccagact cctacgggag gcagcagtgg
 361 ggaatattgc acaatgggcg caagcctgat gcagccatgc cgcgtgtatg aagaaggcct
 421 tcgggttgta aagtactttc agcggggagg aagggagtaa agttaatacc tttgctcatt
 481 gacgttaccc gcagaagaag caccggctaa ctccgtgcca gcagccgcgg taatacggag
 541 ggtgcaagcg ttaatcggaa ttactgggcg taaagcgcac gcaggcggtt tgttaagtca
 601 gatgtgaaat ccccgggctc aacctgggaa ctgcatctga tactggcaag cttgagtctc
 661 gtagaggggg gtagaattcc aggtgtagcg gtgaaatgcg tagagatctg gaggaatacc
 721 ggtggcgaag gcggccccct ggacgaagac tgacgctcag gtgcgaaagc gtggggagca
 781 aacaggatta gataccctgg tagtccacgc cgtaaacgat gtcgacttgg aggttgtgcc
 841 cttgaggcgt ggcttccgga gctaacgcgt taagtcgacc gcctggggag tacggccgca
 901 aggttaaaac tcaaatgaat tgacgggggc ccgcacaagc ggtggagcat gtggtttaat
 961 tcgatgcaac gcgaagaacc ttacctggtc ttgacatcca cagaactttc cagagatgga
1021 ttggtgcctt cgggaactgt gagacaggtg ctgcatggct gtcgtcagct cgtgttgtga
```

```
1081 aatgttgggt taagtcccgc aacgagcgca acccttatct tttgttgcca gcggtccggc
1141 cgggaactca aaggagactg ccagtgataa actggaggaa ggtggggatg acgtcaagtc
1201 atcatggccc ttacgaccag ggctacacac gtgctacaat ggcgcataca aagagaagcg
1261 acctcgcgag agcaagcgga cctcataaag tgcgtcgtag tccggattgg agtctgcaac
1321 tcgactccat gaagtcggaa tcgctagtaa tcgtggatca gaatgccacg gtgaatacgt
1381 tcccgggcct tgtacacacc gcccgtcaca ccatgggagt gggttgcaaa agaagtaggt
1441 agcttaacct tcgggagggc gcttaccact ttgtgattca tgactggggt gaagtcgtaa
1501 caaggtaacc gtaggggaac ctgcggttgg atcacctcct ta
```

There are very few differences between the genes. The Salmonella sequence has three extra bases at the end, and the E-coli gene one extra in the beginning, there are 37 differences within the genes, and no insertions or deletions. Such a comparison is easy to perform in the example above, but if there are many sequences and lots of insertions and deletions, it can be non-trivial to identify the relationships among individual bases. This *multiple sequence alignment problem* is a major problem in comparative genomics, and is discussed in Section 3. In fact, one of our main points is that finding a good multiple alignment is the essence of the ancestral reconstruction problem.

The degree of conservation of the 16S rRNA gene throughout the tree of life means it is a good starting point for reconstructing ancestral genomes. In particular, we can begin modestly by asking only for the ancestral 16S rRNA gene sequences. Such a reconstruction entails the following:

(1) Identifying the 16S gene in many organisms, and determining the sequences.
(2) Obtaining a likely phylogenetic tree that relates the sequences.
(3) Inferring ancestral sequences corresponding to internal nodes in the tree.

This comparative genomics programme was outlined by Woese et al. [**58**]. It followed on the heels of the first comparative genomics papers, written by Linus Pauling and Emile Zuckerkandl [**41, 61, 62**]. They applied fingerprinting techniques to compare amino acid sequences of hemoglobins, finding that distant species have more divergent sequences than related species. The biological problem of identifying the 16S gene and rapidly finding its sequence was solved in [**30**]. More recently, new approaches have been suggested for obtaining 16S rRNA sequences, even from unculturable bacteria, using "community sequencing" approaches [**11**].

The narrow focus of comparative genomics on 16S rRNA ended with the arrival of fast and cheap sequencing technologies. A recent flood of ideas and research inspired by vast amounts of genome sequences has led to the reconstruction of numerous protein sequences and even megabases of boreoeutheriean ancestral chromosomes. Interesting examples of the former are [**24, 28, 54**] and of the latter [**4, 32**]. There have been a number of recent surveys on ancestral reconstruction [**45, 53**], and a book on the topic will appear next year [**31**].

We begin in Section 2 by discussing the "easy" case of ancestral sequence reconstruction, where the phylogenetic tree is known and the alignment of the sequences is trivial. Even in this simplest case, the choice of an effective statistical model for evolution is non-trivial and extremely important. We introduce the reader to these issues by way of the simplest example possible, and provide pointers for further reading. In Section 3 we discuss complexities that arise when inferences need to be made about insertions and deletions, and when the alignment of the sequences is non-trivial. The difficulty of alignment is explored in more detail in

Section 4, where we present evidence, based on our own recent work, that suggests the amount of insertion and deletion in genomes has been vastly underestimated. This has major implications for ancestral genome reconstruction. In Section 5 we discuss the problem of tree reconstruction, which needs to be solved in the case where the phylogenetic history of the genomes being compared is unknown. This leads us to the field of phylogenetics [**49**], where we restrict ourselves to mentioning a number of recent theoretical advances pertinent to ancestral genome reconstruction. We conclude in Section 6 with a list of open problems and a discussion of the role of mathematics, statistics, and computer science in reconstructing ancestral genomes.

2. Reconstructing ancestral sequences: the "easy" case

In the introduction, we mentioned the example of 16S rRNA sequences, and observed that these genes are conserved in all organisms. However, within restricted domains of the tree of life, there are examples of functional elements exhibiting even more sequence conservation than rRNA genes. The term *ultra-conserved elements* was introduced in [**3**], and is used to described genome sequences that have remained unchanged over millions of years. Such sequences were first discovered in vertebrates, and their degree of conservation is astounding.

EXAMPLE 2.1. Consider the sequence

(2.1)   tttaattgaaagaagttaattgaatgaaaatgatcaactaag

It is located in the human genome on chromosome 7, coordinates 156,694,482–156,694,523 (version March 2006). The *identical* sequence appears in the genomes of every other sequenced vertebrate species to date: the chimpanzee, rhesus macaque, cat, dog, cow, mouse, rat, rabbit, armadillo, opossum, chicken, frog, zebrafish, pufferfish and fugufish. An alignment of the sequence (including an extra 6 bases on each end) is shown below:

```
     Human  tatctatttaattgaaagaagttaattgaatgaaaatgatcaactaagcttgta
     Chimp  tatctatttaattgaaagaagttaattgaatgaaaatgatcaactaagcttgta
   Macaque  tatctatttaattgaaagaagttaattgaatgaaaatgatcaactaagcttgta
       Cow  tatctatttaattgaaagaagttaattgaatgaaaatgatcaactaagcttgta
     Mouse  tatctgtttaattgaaagaagttaattgaatgaaaatgatcaactaagcttgta
       Rat  tatctgtttaattgaaagaagttaattgaatgaaaatgatcaactaagtttgta
    Rabbit  tatctatttaattgaaagaagttaattgaatgaaaatgatcaactaagcttgta
       Cat  tatctatttaattgaaagaagttaattgaatgaaaatgatcaactaagcttgta
       Dog  tatctatttaattgaaagaagttaattgaatgaaaatgatcaactaagcttgta
 Armadillo  tatctatttaattgaaagaagttaattgaatgaaaatgatcaactaagcttgta
   Opossum  tatctatttaattgaaagaagttaattgaatgaaaatgatcaactaagcttgta
   Chicken  tatctatttaattgaaagaagttaattgaatgaaaatgatcaactaagcttgta
      Frog  tatctatttaattgaaagaagttaattgaatgaaaatgatcaactaagcttgta
Pufferfish  tatctatttaattgaaagaagttaattgaatgaaaatgatcaactaagcttgta
 Zebrafish  tatctatttaattgaaagaagttaattgaatgaaaatgatcaactaagcttgta
  Fugufish  tatctatttaattgaaagaagttaattgaatgaaaatgatcaactaagcttgta
            ***** ****************************************** *****
```

We postpone providing a precise definition of alignment until Section 3, but remark that for our purposes here it suffices to consider an alignment to be an ordered collection of columns. Each column contains bases from different species

that are *homologous*, i.e., that are derived from a shared common ancestral base. The sequence in (2.1) is called ultra-conserved, because there appear to have been no insertions, deletions or mutations since the common ancestor (* characters indicate columns with all characters identical) among each group of homologous nucleotides. In fact, in [**40**] we prove

THEOREM 2.2. *The probability that the sequence (2.1) was not present in the genome of the ancestor of all vertebrates is less than 10^{-50}, assuming a Jukes-Cantor model of evolution for the sequences.*

The Jukes-Cantor model is a statistical model for the evolution of characters on trees, which is explained below. While the model has many drawbacks and does not describe the full extent and structure of mutation, the tiny probability is robust to changes in the model, and it is fairly certain that (2.1) is the ancestral sequence.

The starting point for specifying an evolutionary model for biological sequences is the data: a collection of k sequences $\sigma^1, \ldots, \sigma^k$ of lengths $n_1, \ldots, n_k$, each with characters from a finite alphabet Σ. We use the notation σ_i^a to denote the ith element of a sequence and by a set of characters $S = \{\sigma^1, \ldots, \sigma^k\}$ we mean the set of $n_1 + n_2 \cdots + n_k$ sequence characters that form the sequences $\sigma^1, \ldots, \sigma^k$. These sequences may be DNA sequences (in which case the alphabet has size 4), amino acid sequences (in which case the alphabet has size 20), or they may be sequences derived from organisms. For example, the alphabet may have size 2, and the sequences may represent the presence or lack of genes, or morphological features in different species. The elements of the sequences are called *characters*, and it is important to note that in problems of interest, k is small but n is usually fairly large (in the case of DNA sequences, n may be in the billions).

We restrict our discussion here to the evolutionary model called the *Cavender–Farris model* [**10**]. This is the simplest of a class of continuous time Markov models for trees that are used in biological sequence analysis, and although it is very simple, it captures many of the elements of the evolutionary models that are used in practice. The Cavender–Farris model is an evolutionary model for *binary* sequences (i.e., the alphabet has size 2), and for sequences of the same length ($n_1 = \cdots = n_k = n$).

The first ingredient of the Cavender-Farris model is a directed rooted tree on k leaves. The root of the tree has biological significance: it is the common ancestor of all the sequences. The leaves of the tree also have special significance: they correspond to the k different sequences whose evolution is being modeled. The internal vertices of the tree correspond to speciation events. Edges of the tree are directed from the root to the leaves, and the directions of the arrows specify the direction of time.

The second ingredient of the Cavender-Farris model is a collection of 2×2 matrices associated with each edge of the tree. These matrices are all of the form

$$(2.2) \qquad \theta_i \quad = \quad \begin{pmatrix} \mu_i & \pi_i \\ \pi_i & \mu_i \end{pmatrix}.$$

where i is an edge of the tree, $0 \leq \pi_i \leq 1$, and $\mu_i = 1 - \pi_i$. These matrices have a biological interpretation: they describe the probabilities of characters changing along branches of the tree. Although in principle these probabilities may be different for each edge *and* for each character within each sequence, the Cavender–Farris model specifies a single probability of change π_i for each edge i. If the vertices

adjacent to the edge i are v_i and w_i with i oriented from v_i to w_i, then for the jth character of the sequence at v_i, π_i is the probability that the jth character of the sequence at w_i is different.

The Cavender-Farris model is an example of a *continuous time Markov chain model on a tree.* To see the connection to Markov models, consider the 2×2 square matrix

$$(2.3) \qquad Q \;=\; \begin{pmatrix} -\alpha & \alpha \\ \alpha & -\alpha \end{pmatrix}, \, \alpha \geq 0.$$

The rows and columns of Q are indexed by $\Sigma = \{0, 1\}$. Note that the matrix Q has the following properties:

$$q_{ij} \geq 0 \quad \text{for} \quad i \neq j,$$

$$\sum_{j \in \Sigma} q_{ij} = 0 \quad \text{for all} \ \ i \in \Sigma,$$

$$q_{ii} < 0 \quad \text{for all} \ \ i \in \Sigma.$$

Any square matrix Q (of arbitrary size) with the above properties is called a *rate matrix.* The following is a straightforward result about continuous time Markov chains [**39**].

THEOREM 2.3. *Let Q be any rate matrix and $\theta(t) = e^{Qt} = \sum_{i=0}^{\infty} \frac{1}{i!} Q^i t^i$. Then*
  (1) $\theta(s + t) = \theta(s) \cdot \theta(t)$ *(Chapman–Kolmogorov equations),*
  (2) $\theta(t)$ *is the unique solution to the forward differential equation*
      $\theta'(t) = \theta(t) \cdot Q$, $\theta(0) = \mathbf{1}$ *for $t \geq 0$ (here $\mathbf{1}$ is the identity matrix),*
  (3) $\theta(t)$ *is the unique solution to the backward differential equation*
      $\theta'(t) = Q \cdot \theta(t)$, $\theta(0) = \mathbf{1}$ *for $t \geq 0$,*
  (4) $\theta^{(k)}(0) = Q^k$.
Furthermore, a matrix Q is a rate matrix if and only if the matrix $\theta(t) = e^{Qt}$ is a stochastic matrix (non-negative with row sums equal to one) for every $t \geq 0$.

Note that for the binary rate matrix (2.3), we have

$$\theta(t) \;=\; \frac{1}{2} \begin{pmatrix} 1 + e^{-2\alpha t_i} & 1 - e^{-2\alpha t_i} \\ 1 - e^{-2\alpha t_i} & 1 + e^{-2\alpha t_i} \end{pmatrix}.$$

The expected number of mutations over time t is the quantity

$$(2.4) \qquad \alpha t \;=\; -\frac{1}{2} \cdot \text{trace}(Q) \cdot t \;=\; -\frac{1}{2} \cdot \log \det\big(\theta(t)\big).$$

This number is called the *branch length.* It can be computed from the substitution matrix $\theta(t)$ and is the expected number of mutations. For this reason it is used instead of t to label edges in a phylogenetic tree.

The matrices θ_i in the Cavender-Farris model are parameterized by α and t:

$$(2.5) \qquad \mu_i = \frac{1}{2}(1 + e^{-2\alpha t}), \quad \pi_i = \frac{1}{2}(1 - e^{-2\alpha t}).$$

One way to specify an evolutionary model is to give a phylogenetic tree T together with Q and an initial distribution for the root of T (which we here assume to be the uniform distribution on Σ). The branch lengths of the edges are unknown parameters, and the objective is to estimate these branch lengths from data. Thus, if the tree T has r edges, then such a model has r free parameters.

Returning to Theorem 2.2, we note that the Jukes-Cantor model is just the Cavender-Farris model with $|\Sigma| = 4$. That is, the Q matrix is given by

$$(2.6) \qquad Q \quad = \quad \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}, \alpha \geq 0.$$

In this case the branch length is given by $3\alpha t$ (this should be compared with (2.4)).

The Cavender-Farris/Jukes-Cantor models are too simple to be used in practice. Point mutations in genome display various asymmetries, and the general reversible Markov model is preferred. The model in (2.7,2.8) is realistic, and was estimated from observed synonymous substitutions (those that do not change the amino acid) in human-mouse-rat alignments [60]. Note that π_a, π_C, π_G and π_T are the equilibrium frequencies, and are also parameters in the model.

$$(2.7) \qquad Q \quad = \quad \begin{pmatrix} -1.05 & 0.19 & 0.71 & 0.15 \\ 0.17 & -0.96 & 0.18 & 0.61 \\ 0.60 & 0.17 & -0.95 & 0.17 \\ 0.15 & 0.72 & 0.20 & -1.07 \end{pmatrix},$$

$$(2.8) \qquad \pi_A = \pi_T = 0.23, \pi_G = \pi_C = 0.27.$$

The models used can be even more general, including local dependencies between sites and different functional categories for the ancestral sequences that alter mutation rates. There is an extensive literature on this topic, as well as many papers discussing the reliability of reconstructed characters in aligned sequences (e.g. [43]).

The final aspect of Theorem 2.2 that we have not yet discussed is a computational one, namely how to compute the probability given the sequences, the tree and the model. The algorithm is known as Felsenstein's algorithm [20] and involves dynamic programming on a tree. In the context of ancestral reconstruction, [44] show how to modify Felsenstein's algorithm for fast joint reconstruction of all ancestral sequences. It is best to view all these algorithms as a special case of the Junction Tree algorithm for a graphical model where the graph is a tree.

We conclude by noting that there is a direct connection between evolutionary models such as the Cavender-Farris model and the emerging field of *algebraic statistics*. This is because the families of probability distributions on the leaves are essentially parameterized algebraic varieties, and for this reason the tools of commutative algebra and algebraic geometry can be used to study the model and develop inference methods. We refer the interested reader to the recent book [39] for an introduction to the subject.

3. Alignment

The models for ancestral reconstruction in Section 2 do not account for insertions and deletions (indels). This is a serious drawback because the homology between muliple sequences is complicated by insertions, deletions, rearrangements, segmental duplications, and other evolutionary events. We restrict our discussion in this section to the issues regarding ancestral reconstruction in the presence of insertions and deletions only.

DEFINITION 3.1. A *partial global multiple alignment* of sequence characters $S = \{\sigma^1, \ldots, \sigma^k\}$ is a partially ordered set $P = \{c_1, \ldots, c_m\}$ together with a surjective function $\varphi : S \to P$ such that $\varphi(\sigma_i^a) < \varphi(\sigma_j^a)$ if $i < j$.

The elements of P correspond to columns of the multiple alignment, and the partial order specifies the order in which columns must appear. Note that the function φ specifies the homology relationships among the sequences: two characters that are mapped to the same element in the poset are homologous. Thus, an alignment is just a pair (P, φ). We call P an *alignment poset*, and note that unless P is a total order (chain), there are columns of the partial multiple alignment whose order is unspecified. A *linear extension* of a partially ordered set $P = \{c_1, \ldots, c_m\}$
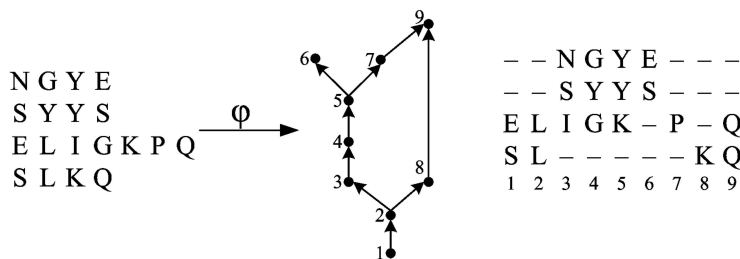

FIGURE 1. A set of four sequences, an alignment poset together with a linear extension, and a global multiple alignment. The function from the set of sequence elements to the alignment poset that specifies the multiple alignment is not shown, but is fully specified by the diagram on the right. For example, the second element in the first sequence is $\sigma_2^1 = G$, and $\varphi(\sigma_2^1)$ corresponds to the fourth column of the multiple alignment.

is a permutation of the elements $c_1, \ldots, c_m$ such that whenever $c_i < c_j$, $i < j$. A *global multiple alignment* is a partial global multiple alignment together with a linear extension of the alignment poset P (see Figure 1). The alignment in example 2.1 is a global multiple alignment where the poset is a chain.

The problem of finding a multiple alignment for a collection of sequences is to find a "good" poset and function φ that are likely to determine the homology correctly. This is a non-trivial problem, and its solution is essential for accurate ancestral reconstruction of sequences. A complete discussion of multiple alignment is beyond the scope of this section, but we briefly review pairwise sequence alignment.

In the case of two sequences, an alignment poset is always a disjoint union of chains. Note that for two sequences of length n and m, there are $\binom{n+m}{n}$ alignments. We consider a simple criterion for selecting an alignment. Given three parameters X, S and M, we assign a score to an alignment as follows:

$$(3.1) \qquad \text{score}(P, \varphi) \quad = \quad M \cdot \#M + X \cdot \#X + S \cdot \#S,$$

where $\#S = |\{x \in P : |\varphi^{-1}(x)| = 1\}|$ (the number of spaces in the alignment), $\#M = |\{a \in \sigma^i, b \in \sigma^j : \varphi(a) = \varphi(b), \text{char } a = \text{char } b\}|$ (the number of matches), and $\#X = |\{a \in \sigma^i, b \in \sigma^j : \varphi(a) = \varphi(b), \text{char } a \neq \text{char } b\}|$ (the number of mismatches). The problem of maximizing (3.1) can be solved efficiently using dynamic programming, using what is known as the *Needleman-Wunsch algorithm* [**36**]. As

with Felsenstein's algorithm, this is just a special case of the Junction Tree algorithm, in this case for a graphical model known as the *pair hidden Markov model*. Despite the name, pair hidden Markov models are non-trivial modifications of standard hidden Markov models models, in that the structure of the models is not fixed. In the graphical model literature, these types of models are referred to as Bayesian multinets. The connection between alignments and pair HMMs provides a probabilistic interpretation for the score in (3.1). For details see [1] or [39].

An alignment specifies the homology relationships among nucleotides or amino acids, but provides no information about ancestral sequences. In order to make inferences about ancestral sequences, it is necessary to use the alignment to estimate the locations, sizes, and times of insertions and deletions.

FIGURE 2. Alignment of 17 species from the CFTR region [50].

EXAMPLE 3.2. Figure 2 shows an alignment of 17 species from the cystic fibrosis transmembrane conductance regulatory region (CFTR) produced for the ENCODE consortium [18]. The phylogenetic tree on the left can be used in conjunction with the alignment to identify indel events. The boxes shown in the figure show the most parsimonious explanation (minimum number of events), based on the tree [50].

Given a phylogenetic tree T, together with a pair hidden Markov model associated to each edge of T, the *tree alignment problem* is to find sequences σ^i where i ranges over all the internal nodes of T such that $\prod_{e=(i,j)\in E(T)} P(\sigma^i, \sigma^j)$ is maximized. The inference performed in Example 3.2 [50] is based on a restricted version of tree alignment. In general, there is a ratio-two approximation algorithm for the problem that runs in quadratic time, together with a polynomial time approximation scheme [56].

4. The limits of ancestral reconstruction: indel saturation

Studies of the reliability of ancestral reconstruction have mostly focused on point mutations and the effects of using different types of evolutionary models [57]. Such studies implicitly assume that multiple alignment, while difficult, is a tractable problem and that indels can be effectively ignored, or else accounted for using procedures such as tree alignment. In this section we provide indirect evidence that even over relatively short time-scales, large number of insertions and deletions make it impossible to align sequences (as there is little or no homology),

and therefore it is impossible to reconstruct ancestral sequences. Our arguments are based on estimating the overall numbers of indels from length differences of homologous segments.

We illustrate our ideas with a calculation based on the comparison of introns in four species of the Drosophila (fruit fly): melanogaster, pseudoobscura, yakuba and virils. Introns are sequences within genes that are spliced out after transcription but before translation, so that they do not contribute to the amino acids in the protein. Although it has been shown that introns can be important for regulating the expression of genes, they consist of mostly non-functional sequence. The reason to examine Drosophila is that multiple genomes have been sequenced during the past year. The reason to look at introns is that it is easy to identify homologous introns when the genes they reside in are unambiguously homologous.

Our methods provide an overall assessment of the number of inserted and deleted nucleotides along each branch of the tree relating the species. The inference is based on a maximum likelihood model for Brownian motion on a tree, originally developed by Felsenstein [21] for modeling changes in gene frequencies over time. Our application of this model to estimating indel rates from length differences of introns is new, and is the first such analysis with more than a pair of species. Previous work on the comparison of lengths of pairs of homologous introns [37, 59] has mainly revealed that there is a correlation between the length differences and the distance between species.

We explain the Gaussian model we consider completely but briefly, for more detail and background we refer the reader to [21, 38]. Let T be an unrooted tree with n leaves, together with labels $\mathbf{v} = v_1, \ldots, v_{2n-3}$ on the edges parameterizing variances of normal distributions. Let x_{ij}^s be quantitative characters generated from a Brownian motion model on a tree. We use the term quantitative character to refer to uncountable state spaces, in this case $x_{ij}^s \in \mathbb{R}$. The ijth contrast is $\Delta x_{ij}^s = x_i^s - x_j^s$. Note that Δx_{ij}^s is drawn from a normal distribution with mean 0 and variance v_{ij}. Let $C = \{i_1 j_1, \ldots, i_{n-1} j_{n-1}\} \subset \binom{[n]}{2}$ be a spanning tree of the complete graph K_n. In the case of four taxa which we will consider (see Figure 3), the C-covariance matrix for $C = \{12, 13, 14\}$ is

$$(4.1) \qquad \Sigma_{12,13,14} = \begin{pmatrix} v_{15} + v_{25} & v_{15} & v_{15} \\ v_{15} & v_{15} + v_{36} + v_{56} & v_{15} + v_{56} \\ v_{15} & v_{15} + v_{56} & v_{15} + v_{46} + v_{56} \end{pmatrix}.$$

We let $\Delta \mathbf{x}_C^s = [\Delta x_{i_1 j_1}^s, \ldots, \Delta x_{i_{n-1} j_{n-1}}^s]$ and denote the determinant of Σ_C by $|\Sigma_C|$. The log-likelihood of the data is given by

$$(4.2) \quad \ln \tilde{L}(\Delta \mathbf{x}_{\binom{[n]}{2}}^s | \mathbf{v}) = -\frac{(n-1)p}{2} \ln 2\pi - \frac{p}{2} \ln |\Sigma_C| + \frac{1}{2} \sum_{s=1}^{p} \left((\Delta \mathbf{x}_C^s)^T \Sigma_C^{-1} \Delta \mathbf{x}_C^s \right).$$

This specifies the model we will use completely. The next Lemma follows from the "Pulley principle" [21].

LEMMA 4.1. *The log-likelihood (4.2) does not depend on the choice of C. Furthermore, (4.2) is linear in the $\binom{n}{2}$ numbers $d_{ij} = \sum_{s=1}^{p} \left(\Delta x_{ij}^s \right)^2$.*

In other words, if we let $\mathbf{d} = \{d_{ij}\}_{i,j=1}^n$, then the maximum likelihood estimator

$$(4.3) \qquad\qquad\qquad \widehat{\mathbf{v}_d} = \text{argmax}_{\mathbf{v}} \quad \ln L(\mathbf{d}|\mathbf{v}).$$
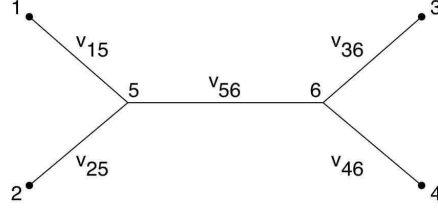
FIGURE 3. Tree with four leaves.

is well-defined and it therefore makes sense to refer to the "log-likelihood function for the contrast Brownian motion model on a tree".

EXAMPLE 4.2 (n=3, from [21]). The likelihood equation becomes

$$\ln L(d_{12}, d_{13,}, d_{23}|v_{14}, v_{24}, v_{34}) = -p\ln 2\pi - \frac{p}{2}\ln(v_{14}v_{24} + v_{14}v_{34} + v_{24}v_{34})$$
$$-\frac{v_{14}d_{23} + v_{24}d_{13} + v_{34}d_{12}}{2(v_{14}v_{24} + v_{14}v_{34} + v_{24}v_{34})}.$$

The critical equations are easy to solve and one finds that

$$\widehat{v_{14}} = (d_{12} + d_{13} - d_{23})/(2p),$$
$$\widehat{v_{24}} = (d_{23} + d_{12} - d_{13})/(2p),$$
$$\widehat{v_{34}} = (d_{13} + d_{23} - d_{12})/(2p).$$

Note that in the case $p = 1$, if $\mathbf{d}$ is a tree metric then $\widehat{V_{ij}} = d_{ij}$. This is true for general n, and is a restatement of the fact that the maximum likelihood estimator is consistent.

EXAMPLE 4.3 (n=4). There are five critical equations in the $n = 4$ case:

$$v_{25}(v_{36} + v_{46}) + v_{56}(v_{36} + v_{46}) = (d_{34}(v_{25} + v_{56}) + d_{24}v_{35} + d_{23}v_{45})/(2p),$$
$$v_{15}(v_{36} + v_{46}) + v_{56}(v_{36} + v_{46}) = (d_{34}(v_{15} + v_{56}) + d_{13}v_{46} + d_{14}v_{36})/(2p),$$
$$v_{46}(v_{36} + v_{46}) + v_{56}(v_{14} + v_{24}) = (d_{12}(v_{46} + v_{56}) + d_{14}v_{25} + d_{24}v_{15})/(2p),$$
$$v_{36}(v_{36} + v_{46}) + v_{56}(v_{14} + v_{24}) = (d_{12}(v_{36} + v_{56}) + d_{23}v_{15} + d_{13}v_{25})/(2p),$$
$$(v_{15} + v_{25})(v_{36} + v_{46}) = (d_{34}(v_{15} + v_{25}) + d_{12}(v_{36} + v_{46}))/(2p).$$

The solution of these equations is an exercise in elimination, and can be done using Gröbner bases methods [39]. Using the pulley principle we can restrict ourselves to $\widehat{v_{15}} + \widehat{v_{25}} = d_{12}$ and $\widehat{v_{36}} + \widehat{v_{46}} = d_{34}$, in which case we find one unique critical point: the global maximum. The solution consists of 5 enormous rational functions in the d_{ij} which we omit here due to lack of space.

The data analyzed was obtained from [59]. We restricted our attention to introns in the range of $100 - 500$bp, and cleaned up the dataset by removing duplicates or cases with ambiguous homology. We then computed, for each pair of species, the "distance"

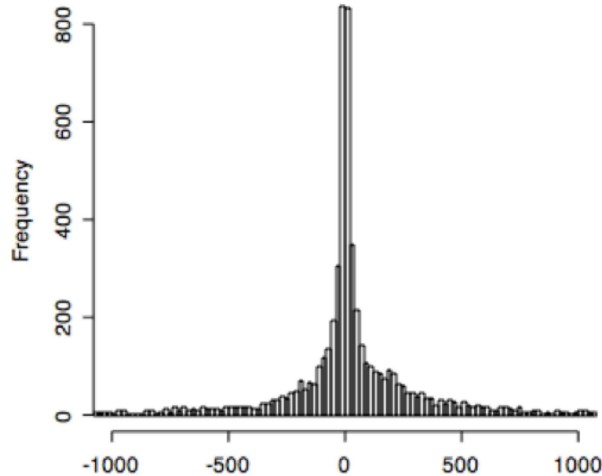$$(4.4) \qquad\qquad d_{ij} = \sum_{s=1}^{p}(\Delta x_{ij}^{s})^2,$$

FIGURE 4. Differences in lengths between homologous introns in Drosophila melanogaster and pseudoobscura. Introns with similar numbers of insertions and deletions display little length difference, but that does not mean that introns with small length differences have not undergone insertion and deletion.

where Δx_{ij}^s is the difference in length between the sth pair of introns in species i and j. This is just the quantity that appears in Lemma 4.1. It is important to note that the use of the Brownian motion model on the d_{ij} constitutes an approximation to a Poisson process model for indels. We omit the details of this relationship and again refer to [**38**].

Returning to the data, we show a histogram of the differences between the intron lengths for a pair of species in Figure 4. These are the raw data we use to compute the d_{ij} distances, and then the maximum likelihood estimates for the branch lengths, which are the total number of indels. The maximum likelihood estimates are

$$
\begin{aligned}
\widehat{v}_{dmel,5} &= 3204, \\
\widehat{v}_{dyak,5} &= 2521, \\
\widehat{v}_{dpse,6} &= 9595, \\
\widehat{v}_{dvir,6} &= 49894, \\
\widehat{v}_{56} &= 16169,
\end{aligned}
$$

where vertices $5, 6$ are the internal vertices in the four taxa tree (as in Figure 3). In other words, we estimate that the total number of inserted and deleted bases between D. melanogaster and D. Yakuba is 5725, and between D. melanogaster and D. pseudoobscura we obtain $3204 + 16169 + 9595 = 28968$.

These numbers are much larger than the mean length of the introns we considered (note that the longest introns had length 500). The conclusion is that the *total* number of inserted and deleted bases is far larger than the size of the intron. This may seem surprising at first, but is a reflection of the fact that insertions and

deletions cancel each other out, and therefore a small difference in intron length does not necessarily indicate a lack of indel activity. In fact, the intuition that homologous introns of similar length must contain homologous nucleotides is false. The difference of two Poisson distributions is Skellam distributed, and if the rates are the same then the peak is at 0, exactly what we see in Figure 4.

The Brownian motion model we have proposed is (too) simple; it is the indel equivalent of the Jukes-Cantor model for point mutations, but it is sufficiently realistic to suggest that it may be impossible to reconstruct ancestral intron due to excessive indel turnover. The large amount of insertion and deletion should not be surprising in the light of the existence of *transposable elements*. These are repetitive elements that make up a large fraction of many genomes. The term transposable elements groups several subclasses of elements that replicate autonomously in the genome, either through reverse transcription, or directly from DNA to DNA via excision and repair. Up to half of the human genome is composed of such elements, and although they are sometimes thought of as "parasitic elements", somewhat like viruses, they clearly play an integral role in shaping genome evolution, and in many cases are believed to influence gene function. Unfortunately, they confound attempts at reconstructing genomes, by virtue of creating enormous turnover in the sequences. At the very least, a complete catalog of such elements will be essential for reconstructing ancestral genomes.

5. Tree reconstruction

In the previous sections we have been assuming that the phylogenetic tree for the species under consideration is known. This assumption is, unfortunately, rarely justifiable. Molecular based phylogenies may not conclusively determine certain branchings in a tree, and fossil-based phylogenies tend to have low resolution. We mention two best-case examples where despite substantial work there is still some disagreement as to the actual phylogeny. In vertebrates, molecular techniques are not in agreement with other methods used for the rodents (the so-called "rodent problem" [**2, 52**]), and in Drosophila there is disagreement about the splits among Drosophila erecta, yakuba and melanogaster [**42**]. In other branches of the tree of life, the situation can be that nothing at all is known about the details of the phylogeny. Thus, phylogenetic trees must be inferred, and the topology of the trees has a direct bearing on the reconstructed ancestral sequences [**46**].

We begin by describing a likelihood-based strategy that can be followed, but that is computationally infeasible in practice: for each tree, the probability of the known sequences may be computed for a specific evolutionary model, and one can select the tree/evolutionary model combination with maximal likelihood. This is known as the *maximum likelihood approach* to phylogeny reconstruction. The reason the algorithm proposed above is computationally intractable for trees with many leaves is that the number of binary trees with k leaves is $(2k-5)!!$. Moreover, the problem of finding the maximum likelihood branch lengths for a fixed tree is very difficult [**25**]. The field of phylogenetics research is very active and it is only recently that the following result was published, quantifying the difficulty of the tree reconstruction problem:

THEOREM 5.1 ([**12**]). *Given a set of binary strings, all of the same length, and a negative number L, it is NP-hard to determine whether there is a tree T such*

that the log likelihood of the sequences for the tree T with optimal branch lengths is greater than L.

On the positive side, there are theoretically sound approaches to phylogenetic reconstruction that are also practical for large datasets. In the context of ancestral reconstruction there are often many taxa to be considered, and the favored approach is *neighbor joining* [**47**] (sometimes other closely related distance-based algorithms are used [**8**]). The neighbor joining algorithm takes as input a *dissimilarity map* on a set of taxa X. This is a map $\delta : X \times X \to \mathbb{R}$ that satisfies $\delta(i, j) = \delta(j, i)$ and $\delta(i, i) = 0$. The quantities $\delta(i, j)$ are maximum likelihood estimates of the branch length (see 2.4) between every pair of taxa. The algorithm is:

(1) Given a dissimilarity map δ, compute the *Q-criterion*

$$Q_\delta(i, j) = (n - 2)\delta(i, j) - \sum_{k \neq i} \delta(i, k) - \sum_{k \neq j} \delta(j, k).$$

Then select a pair a, b that minimize Q_δ as motivated by the following theorem:

THEOREM 5.2 ([**47**]). *Let δ_T be the tree metric corresponding to the tree T. The pair a, b that minimizes $Q_{\delta_T}(i, j)$ is a cherry in the tree.*

(2) If there are more than three taxa, replace the putative cherry a and b with a leaf j_{ab}, and construct a new dissimilarity map where $\delta(i, j_{ab}) = \frac{1}{2}(\delta(i, a) + \delta(i, b))$. This is called the *reduction step*.
(3) Repeat (1) and (2) until there are three taxa.

Neighbor-joining is fast: existing implementations run in $O(n^3)$ where n is the number of taxa, and it has been observed (empirically) to produce good results [**29**]. However, despite the ubiquitous use of the algorithm, very little about it has been understood until recently. Exciting new results include:

- The development of fast neighbor-joining which achieves an optimal run time of $O(n^2)$ [**17**].
- A uniqueness theorem for the algorithm [**7**].
- An answer to the question of "what does neighbor-joining optimize"? [**23**].
- An answer to the question of "when (and why) does neighbor-joining work"? [**35**].

Together, these results provide new insight into the algorithm, and open up the possibility for significant improvements in accuracy.

Returning to maximum likelihood phylogenetic reconstruction, recent results also show that it is possible to efficiently reconstruct the topology of trees (with high probability) using likelihood models of the type described in Section 2 given only polylogarithmic quantities of data.

THEOREM 5.3 ([**13**]). *Under the Cavender-Farris model, there is a constructive algorithm that reconstructs almost all trees on k leaves with sequences of length $k = O(poly(\log k))$.*

A recent related result [**34**] provides an alternative analysis that quantitatively couples the reconstruction problem with the ancestral reconstruction problem. Indeed, it appears that ancestral sequence reconstruction and tree reconstruction are far more related than originally thought.

6. Open problems and discussion

A recent survey article [**45**] proposes that "an integrated, multi-disciplinary approach is needed in order to make progress on ancestral genome reconstruction". We agree with this point of view, and in the spirit of the proposal offer an invitation to mathematicians, statisticians and computer scientists by highlighting some open problems that may form a starting point for research and collaboration. We focus on problems important for biology, but many of the questions also lead to interesting mathematics [**51**].

The Cavender-Farris model introduced in Section 2 is the simplest example of an evolutionary model. A central problem in genomics is to find appropriate models that effectively capture the mechanisms by which sequences changes, but that are also useful for inference. An important class of models that have been proposed are phylogenetic hidden Markov models [**33, 48**].

Problem 6.1 (Phylogenetic hidden Markov models). Find efficient algorithms for inference with phylogenetic hidden Markov models.

See [**27**] for an introduction to this problem and some first steps exploring the use of variational methods and other graphical model techniques.

In Section 4, we raise the issue of alignability of sequences, and the implications for ancestral genome reconstruction. There are other approaches to addressing the problem of alignability: In [**14**], we show that the choice of parameters is crucial for correctly identifying homologous transcription factor binding sites. The methods used are those of *parametric alignment*, which is a geometric approach to studying the dependence of optimal alignments on parameters. We propose the following problem based on [**22**].

Problem 6.2 (Parametric ancestral reconstruction). Develop polyhedral algorithms for the ancestral reconstruction problem. In particular, what implications does the *Jukes-Cantor function* of [**14**], have for ancestral reconstruction?

In a similar vein, and inspired by our observation in Section 4 that the number of indels in introns may preclude ancestral reconstruction, we ask:

Problem 6.3 (Indel saturation). What are the limits on ancestral reconstruction as determined by indel rates and distances?

In the field of phylogenetics, we offer two problems chosen for their specific relevance to the ancestral genome reconstruction problem. For readers interested in algebraic geometry, we mention [**19**] for further mathematical problems.

Problem 6.4 (Tree reconstruction and alignment). Find efficient algorithms for reconstructing a tree under the tree alignment model.

The next problem is especially important for ancestral reconstruction of bacterial genomes where there is a lot of horizontal transfer:

Problem 6.5 (Consistency theorems for networks). Extend the robustness analysis of [**35**] to generalizations of the neighbor joining algorithm that project dissimilarity maps onto phylogenetic networks [**9, 26**] rather than trees.

The reconstruction of ancestral genomes involves more than the inference of ancestral sequences based on groups of homologous extant nucleotides. The order of sequences is also important, and an important component of *whole* genome

reconstruction is the inference of the ancestral order of genomic segments. The problem is closely related to the whole genome alignment problem [**15**]. In this regard, Definition 3.1 is too restrictive. For example, it does not allow for homology relationships where there have been rearrangements, inversions, and segmental duplications. In our opinion, a major problem that needs to be solved, where a close collaboration between biologists and mathematicians is necessary is:

Problem 6.6 (What is an alignment?). Provide a definition for whole genome alignment that is based on a comprehensive biological definition of homology.

There are formulations of alignment different from 3.1, but they are also too restrictive. Nevertheless, we mention one important approach to inference of ancestral order:

DEFINITION 6.7. A *reversal* alignment between two genomes is a signed permutation.

For example, the signed permutation

$$(6.1) \qquad\qquad 1\,7^-\,6\,10^-\,9\,8^-\,2\,11^-\,3^-\,5\,4$$

is a reversal alignment between the human and mouse X chromosomes (this is an example from [**5**]). This means that there is a division of the human X chromosome into 11 pieces (equivalently 11 breakpoints) such that if they are labeled, in order, $1\,2\,3\,4\,5\,6\,7\,8\,9\,10\,11$, then they appear in the mouse in the order (6.1). Note that the negative signs specify the direction of the segments, with a negative indicating reversal and complementation. A reversal operation involves reversing the order of a segment of a signed permutation, and flipping the signs. For example, by a reversal of (6.1) can consist of changing the segment $11^-\,3^-\,5\,4$ to $4^-\,5^-\,3\,11$. Biologically, reversals correspond to rearrangement events. The reconstruction of ancestral order is equivalent to

Problem 6.8 (The median problem). Given a phylogenetic tree T, a distance measure between signed permutations, and signed permutations labeling the leaves of T, find signed permutations π^i where i ranges over all the internal nodes of T such that the sum of the distances between permutations adjacent in the tree is minimized.

The case where the distance measure is the reversal distance is already interesting and difficult, but in practice more complex distance measures need to be used (that allow for multiple chromosomes and other events, such as duplications). For more on the problem see [**6, 16, 55**].

We conclude by noting that although we have not discussed it in this paper, ancestral reconstructions of proteins can be sequenced and tested for their physiochemical properties (e.g., [**28, 54**]). Thus, ancestral reconstructions are not merely theoretical exercises. This exciting aspect of the field continues to be developed, and will hopefully lead to tests not just of genes, but also of ancestral regulatory elements and larger genome segments.

References

[1] M. Alexandersson, N. Bray and L. Pachter, *Pair hidden Markov models*, special review in the Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics (L.B. Jorde, P. Little, M. Dunn and S. Subramanian, eds.), 2005.

[2] J. Al-Aidroos and S. Snir, *Analysis of point mutations in vertebrate genomes*, in Algebraic Statistics for Computational Biology (L. Pachter and B. Sturmfels eds.), Cambridge University Press, 2006.

[3] G. Bejerano, M. Pheasant, I. Makunun, S. Stephen, W.J. Kent, J.S. Mattick and D. Haussler, *Ultraconserved elements in the human genome*, Science 304:1321–1325, 2004.

[4] M. Blanchette, E.D. Green, W. Miller and D. Haussler, *Reconstructing large regions of an ancestral mammalian genome in silico*, Genome Research 14:2412–2423, 2004.

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

[6] G. Bourque, G. Tesler and P.A. Pevzner, *The convergence of cytogenetics and rearrangement-based models for ancestral genome reconstruction*, Genome Research 16:311–313, 2006.

[7] D. Bryant, *On the uniqueness of the selection criterion in neighbor-joining*, Journal of Classification 22:3–15, 2005.

[8] W. Cai, J. Pei and N.V. Grishin, *Reconstruction of ancestral protein sequences and its applications*, BMC Evolutionary Biology 4:33, 2004.

[9] D. Bryant and V. Moulton, *NeighborNet: An agglomerative method for the construction of phylogenetic networks*, Evolution 21:255–265, 2004.

[10] J. Cavender, *Taxonomy with confidence*, Mathematical Biosciences 40:271–280, 1978.

[11] K. Chen and L. Pachter, *Bioinformatics for whole-genome shotgun sequencing of microbial communities*, PLoS Computational Biology 1:e24, 2005.

[12] B. Chor and T. Tuller, *Maximum likelihood of evolutionary trees is hard*, in Proceedings of RECOMB 2005.

[13] C. Daskalakis, C. Hill, A. Jaffe, R. Miheascu, E. Mossel and S. Rao, *Maximal accurate forests from distance matrices*, in Proceedings of RECOMB 2006.

[14] C. Dewey, P. Huggins, K. Woods, B. Sturmfels and L. Pachter, *Parametric alignment of Drosophila genomes*, PLoS Computational Biology 2:e73, 2006.

[15] C. Dewey and L. Pachter, *Evolution at the nucleotide level: the problem of multiple whole-genome alignment*, Human Molecular Genetics 15:R51–R56, 2006.

[16] R. Durrett and Y. Interian, *Genomic midpoints: computationa and evolution implications*, submitted.

[17] I. Elias and J. Lagergren, *Fast neighbor joining*, Proceedings of the International Colloquium on Automata, Languages and Programming (ICALP), 2005.

[18] The ENCODE Project Consortium, *The ENCODE (ENCyclopedia of DNA Elements) Project*, Science 306:636–640, 2004.

[19] N. Eriksson, K. Ranestad and B. Sturmfels, *Phylogenetic algebraic geometry*, in Projective Varieties with Unexpected Properties, C. Ciliberto et al. (eds), Water de Gruyter, Berlin 2005.

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

[21] J. Felsenstein, *Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates*, Evolution 35:1229–1242, 1981.

[22] D. Fernandez-Baca, B. Venkatachalam, A. Alberto, C. Maxime and P. Kunsoo, *Parametric analysis for ungapped Markov models of evolution*, in Proceedings of the Conference on Combinatorial Pattern Matching, 2005.

[23] O. Gascuel and M. Steel, *Neighbor joining revealed*, Molecular Biology and Evolution 23:1997–2000, 2006.

[24] E.A. Gaucher, J.M. Thomson, M.F. Burgan and S.A. Benner, *Inferring the paleoenvironment of ancient bacteria on the basis of resurrected proteins*, Nature 425:285–288, 2003.

[25] S. Hoşten, A. Khetan and B. Sturmfels, *Solving the likelihood equations*, Foundations of Computational Mathematics 5:389–407, 2005.

[26] T.N.D. Huynh, J. Jansson, N.B. Nguyen and W.-K. Sung, *Constructing a smallest refining galled phylogenetic network*, in Proceedings of RECOMB 2005.

[27] V. Jojic, N. Jojic, C. Meek, D. Geiger, A. Siepel, D. Haussler and D. Heckerman, *Efficient approximations for learning phylogenetic HMM models from data*, Bioinformatics:20:i161–i168.

[28] D. Kuang, Y. Yao, D. MacLean, M. Wang, D.R. Hampson and B.S.W. Chang, *Ancestral reconstruction of the ligand binding pocket of family C G-protein coupled receptors*, Proceedings of the National Academy of Sciences of the USA, 103:14050–14055, 2006.

[29] S. Kumar and S.R. Gadagker, *Efficiency of the neighbor-joining method in reconstructing evolutionary relationships in large phylogenies*, Journal of Molecular Evolution 51:544–553, 2000.

[30] D.J. Lane, B. Pace, G.J. Olsen, D.A. Stahl, M.L. Sogin and M.R. Pace, *Rapid determination of 16S ribosomal RNA sequences for phylogenetics analysis.* Proceedings of the National Academy of Sciences, USA 20:6955-6969, 1985.

[31] D.A. Liberles (editor), *Ancestral Sequence Reconstruction*, Oxford University Press, to appear in 2007.

[32] J. Ma, L. Zhang, B.B. Suh, B.J. Raney, R.C. Burhans, W.J. Kent, M. Blanchette, D. Haussler and W. Miller, *Reconstructing contiguous regions of an ancestral genome*, Genome Research 16:1557–1565, 2006.

[33] J. McAuliffe, L. Pachter and M.I. Jordan, *Multiple sequence functional annotation and the generalized hidden Markov phylogeny*, Bioinformatics 20:1850–1860, 2004.

[34] R.H. Mihaescu, D. Adkins, C. Hill, A. Jaffe and S. Rao, *A simple quadratic time algorithm for accurate phylogeny reconstruction from logarithmic-sized data*, submitted.

[35] R.H. Mihaescu, D. Levy and L. Pachter, *Why neighbor-joining works*, submitted (arXiv cs.DS/0602041).

[36] S.B. Needleman and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins*, Journal of Molecular Biology 48:443–453, 1970.

[37] A.Y. Ogurtsov, S. Sunyaev and A.S. Kondrashov, *Indel-based evolutionary distance and mouse-human divergence*, Genome Research 14:1610–1616, 2004.

[38] L. Pachter, *The majority of divergence between DNA sequences is due to indels*, submitted.

[39] L. Pachter and B. Sturmfels, *Algebraic Statistics for Computational Biology*, Cambridge University Press 2005.

[40] L. Pachter and B. Sturmfels, *The mathematics of phylogenomics.* SIAM review, in press.

[41] L. Pauling and E. Zuckerkandl, *Chemical paleogenetics: molecular restoration studies of extinct forms of life*, Acta Chem. Scand. 17:89, 1963.

[42] D. Pollard, V.N. Iyer, A.M. Moses and M.B.Eisen, *Whole genome phylogeny of the Drosophila melanogaster species subgroup: widespread discordance with species tree & evidence for incomplete lineage sorting*, PLoS Genetics, advance access 2006.

[43] T. Pupko, A. Doron-Faigenboim, D.A. Liberles and G. M. Cannarozzi, *Probabilistic models and their impact on the accuracy of reconstructed ancestral sequences*, in Ancestral Sequence Reconstruction, D.A. Liberles (editor), Oxford University Press, to appear in 2007.

[44] T. Pupko, I. Pe'er, R. Shamir and D. Graur, *A fast algorithm for joint reconstruction of ancestral amino acid sequences*, Molecular Biology and Evolution 17:890–896, 2000.

[45] M. Rocchi, N. Archidiacono and R. Stanyon, *Ancestral genomes reconstruction: an integrated, multi-disciplinary approach is needed*, Genome Research 16:1441–1444, 2006.

[46] H.A. Ross, D.C. Nickle, Y. Liu, L. Heath, M.A. Jensen, A.G. Rodrigo and J.I. Mullins, *Sources of variation in ancestral sequence reconstruction for HIV-1 enevelope genes*, Evolutionary Bioinformatics 2:18–41, 2006.

[47] N. Saitou and M. Nei, *The neighbor joining method: a new method for reconstructing phylogenetic trees*, Molecular Biology and Evolution 4:406–425,1987.

[48] A. Siepel and D. Haussler, *Phylogenetic hidden Markov models*, in Statistical Methods in Molecular Evolution (R. Nielsen editor), Springer, NY, p 325–351, 2005.

[49] C. Semple and M. Steel, *Phylogenetics*, Oxford University Press, 2003.

[50] S. Snir and L. Pachter, *Phylogenetic profiling of insertions and deletions*, in Proceedings of RECOMB 2006.

[51] B. Sturmfels, *Can biology lead to new theorems?*, Annual report of the Clay Mathematics Institute, 2005.

[52] J.W. Thomas et al., *Comparative analysis of multi-species sequences from targeted genomic regions*, Nature 424:788–783, 2003.

[53] J.W. Thornton, *Resurrecting ancient genes: experimental analysis of extinct molecular*, Nature Reviews Genetics 5:366–375, 2004.

[54] J.W. Thornton, E. Need and D. Crews, *Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling*, Science 301:1714–1717, 2003.

[55] L.-S. Wang and T. Warnow, *Reconstructing chromosomal evolution*, SIAM Journal on Computing, in press.

[56] L. Wang and D. Gusfield, *Improved approximation algorithms for tree alignment*, Journal of Algorithms 25:255–273, 1997.

[57] P.D. Williams, D.D. Pollock, B.P. Blackburne and R.A. Goldstein, *Assessing the accuracy of ancestral protein reconstruction methods*, PLoS Computational Biology, advance access (2006).

[58] Woese, C.R. and GE Fox, *Phylogenetic structure of the prokaryotic domain: the primary kingdoms.* Proceedings of the National Academy of Sciences, USA 74:5088–5090.

[59] M. Yandell, C.J. Mungall, C. Smith, S. Prochnik, J. Kamniker, G. Hartzell, S. Lewis and G.M. Rubin, *Large-scale trends in the evolution of gene structures within 11 animal genomes*, PLoS Computational Biology 2:e15, 2006.

[60] V.B. Yap and L. Pachter, *Identification of evolutionary hotspots in the rodent genomes*, Genome Research 14:574–579, 2004.

[61] E. Zuckerkandl and L. Pauling, *Molecular disease, evolution, and genetic heterogeneity*, in M. Kahsa and B. Pullman eds. Horizons in biochemistry. Academic Press, New York. 1962.

[62] E. Zuckerkandl and L. Pauling, *Molecules as documents of evolutionary history.* Journal Theorertical Biology 8:357–366, 1965.

Departments of Mathematics and Computer Science, UC Berkeley.