SIAM Review
Vol. 49, No. 1, pp. 3–31

The Mathematics of Phylogenomics*

Lior Pachter†
Bernd Sturmfels†

Abstract. The grand challenges in biology today are being shaped by powerful high-throughput technologies that have revealed the genomes of many organisms, global expression patterns of genes, and detailed information about variation within populations. We are therefore able to ask, for the first time, fundamental questions about the evolution of genomes, the structure of genes and their regulation, and the connections between genotypes and phenotypes of individuals. The answers to these questions are all predicated on progress in a variety of computational, statistical, and mathematical fields. The rapid growth in the characterization of genomes has led to the advancement of a new discipline called phylogenomics. This discipline results from the combination of two major fields in the life sciences: genomics, i.e., the study of the function and structure of genes and genomes; and molecular phylogenetics, i.e., the study of the hierarchical evolutionary relationships among organisms and their genomes. The objective of this article is to offer mathematicians a first introduction to this emerging field, and to discuss specific mathematical problems and developments arising from phylogenomics.

Key words. genomics, phylogenetics, genetic code, algebraic statistics, hidden Markov model, sequence alignment, ultraconservation

AMS subject classifications. Primary, 92D20; Secondary, 62-02

DOI. 10.1137/050632634

The lack of real contact between mathematics and biology is either a tragedy, a scandal or a challenge, it is hard to decide which.

–Gian-Carlo Rota [34, p. 2]

1. Introduction.

The grand challenges in biology today are being shaped by powerful high-throughput technologies that have revealed the genomes of many organisms, global expression patterns of genes, and detailed information about variation within populations. We are therefore able to ask, for the first time, fundamental questions about the evolution of genomes, the structure of genes and their regulation, and the connections between genotypes and phenotypes of individuals. The answers to these questions are all predicated on progress in a variety of computational, statistical, and mathematical fields [35].

The rapid growth in the characterization of genomes has led to the advancement of a new discipline called *phylogenomics*. This discipline, whose scope and potential was first outlined in [22], results from the combination of two major fields in the

*Received by the editors May 29, 2005; accepted for publication (in revised form) September 30, 2005; published electronically January 30, 2007.
http://www.siam.org/journals/sirev/49-1/63263.html

†Department of Mathematics, University of California, Berkeley, CA (lpachter@math.berkeley.edu, bernd@math.berkeley.edu). The first author was supported by a grant from the NIH (R01-HG2362-3), a Sloan Foundation Research Fellowship, and an NSF CAREER award (CCF-0347992). The second author was supported by the NSF (DMS-0200729, DMS-0456960).

life sciences: *genomics*, i.e., the study of the function and structure of genes and genomes; and *molecular phylogenetics*, i.e., the study of the hierarchical evolutionary relationships among organisms and their genomes. The objective of this article is to offer mathematicians a first introduction to this emerging field, and to discuss specific problems and developments arising from phylogenomics.

The mathematical tools to be highlighted in this paper are statistics, probability, combinatorics, and—last but not least—algebraic geometry. Emphasis is placed on the use of *algebraic statistics*, which is the study of statistical models for discrete data using algebraic methods. See [44, section 1] for details. Several models which are relevant for phylogenomics are shown to be algebraic varieties in certain high-dimensional spaces of probability distributions. This interplay between statistics and algebraic geometry offers a conceptual framework for understanding and developing combinatorial algorithms for biological sequence analysis. It is our hope that this will contribute to some "real contact" between mathematics and molecular biology.

This paper is organized as follows. In section 2 we begin by reviewing the organization and structure of genomes. This section is meant as a brief tutorial, aimed at readers who have a little or no background in molecular biology. It offers definitions of the relevant biological terminology.

Section 3 describes a very simple example of a statistical model for inferring information about the genetic code. The point of this example is to explain the philosophy of algebraic statistics: *model means algebraic variety*.

A more realistic model, which is widely used in computational biology, is the *hidden Markov model* (HMM). In section 4 we explain this model and discuss its applications to the *gene finding* problem. Another key problem is the alignment of biological sequences. Section 5 reviews the statistical models and combinatorial algorithms for *sequence alignment*. We also discuss the relevance of *parametric inference* [43].

In section 6 we present statistical models for the evolution of biological sequences. These models are algebraic varieties associated with phylogenetic trees, and they play a key role in inferring the ancestral relationships among organisms and in identifying regions in genomes that are under selection.

Section 7 gives an introduction to the field of *phylogenetic combinatorics*, which is concerned with the combinatorics and geometry of finite metric spaces and their application to data analysis in the life sciences. We shall discuss the space of all trees [9], the neighbor-joining algorithm for projecting metrics onto this space, and several natural generalizations of these concepts.

In section 8 we go back to the data. We explain how one obtains and studies DNA sequences generated by genome sequencing centers, and we illustrate the mathematical models by estimating the probability that the DNA sequence in Conjecture 1 occurred by chance in ten vertebrate genomes.

2. The Genome. Every living organism has a genome, made up of deoxyribonucleic acids (DNA) arranged in a double helix [61], which encodes (in a way to be made precise) the fundamental ingredients of life. Organisms are divided into two major classes: *eukaryotes* (organisms whose cells contain a nucleus) and *prokaryotes* (for example, bacteria). In our discussion we focus on genomes of eukaryotes and, in particular, the human genome [38, 59].

Eukaryotic genomes are divided into *chromosomes*. The human genome has two copies of each chromosome. There are 23 pairs of chromosomes: 22 *autosomes* (two copies each in both men and women) and two *sex chromosomes*, which are denoted X and Y. Women have two X chromosomes, while men have one X and one Y chromosome. Parents pass on a mosaic of their pair of chromosomes to their children.

Table 2.1 *The genetic code.*

	T	C	A	G
T	TTT ↦ Phe	TCT ↦ Ser	TAT ↦ Tyr	TGT ↦ Cys
	TTC ↦ Phe	TCC ↦ Ser	TAC ↦ Tyr	TGC ↦ Cys
	TTA ↦ Leu	TCA ↦ Ser	TAA ↦ *stop*	TGA ↦ *stop*
	TTG ↦ Leu	TCG ↦ Ser	TAG ↦ *stop*	TGG ↦ Trp
C	CTT ↦ Leu	CCT ↦ Pro	CAT ↦ His	CGT ↦ Arg
	CTC ↦ Leu	CCC ↦ Pro	CAC ↦ His	CGC ↦ Arg
	CTA ↦ Leu	CCA ↦ Pro	CAA ↦ Gln	CGA ↦ Arg
	CTG ↦ Leu	CCG ↦ Pro	CAG ↦ Gln	CGG ↦ Arg
A	ATT ↦ Ile	ACT ↦ Thr	AAT ↦ Asn	AGT ↦ Ser
	ATC ↦ Ile	ACC ↦ Thr	AAC ↦ Asn	AGC ↦ Ser
	ATA ↦ Ile	ACA ↦ Thr	AAA ↦ Lys	AGA ↦ Arg
	ATG ↦ Met	ACG ↦ Thr	AAG ↦ Lys	AGG ↦ Arg
G	GTT ↦ Val	GCT ↦ Ala	GAT ↦ Asp	GGT ↦ Gly
	GTC ↦ Val	GCC ↦ Ala	GAC ↦ Asp	GGC ↦ Gly
	GTA ↦ Val	GCA ↦ Ala	GAA ↦ Glu	GGA ↦ Gly
	GTG ↦ Val	GCG ↦ Ala	GAG ↦ Glu	GGG ↦ Gly

The sequence of DNA molecules in a genome is typically represented as a sequence of letters, partitioned into chromosomes, from the four letter alphabet $\Omega = \{A, C, G, T\}$. These letters correspond to the bases in the double helix, that is, the *nucleotides* adenine, cytosine, guanine, and thymine. Since every base is paired with an opposite base (A with T and C with G in the other half of the double helix), in order to describe a genome it suffices to list the bases in only one strand. However, it is important to note that the two strands have a directionality which is indicated by the numbers $5'$ and $3'$ on the ends (corresponding to carbon atoms in the helix backbone). The convention is to represent DNA in the $5' \to 3'$ direction. The human genome consists of approximately 2.8 billion bases, and has been obtained using high-throughput sequencing technologies that can be used to read the sequence of short DNA fragments hundreds of bases long. *Sequence assembly algorithms* are then used to piece together these fragments [39]. See also [44, section 4].

Despite the tendency to abstract genomes as strings over the alphabet Ω, one must not forget that they are highly structured: for example, certain subsequences within a genome correspond to *genes*. These subsequences play the important role of encoding *proteins*. Proteins are polymers made of twenty different types of amino acids. Within a gene, triplets of DNA, known as *codons*, encode the amino acids for the proteins. This is known as the *genetic code*. Table 2.1 shows the 64 possible codons and the twenty amino acids they code for. Each amino acid is represented by a three letter identifier ("Phe" = Phenylalanine, "Leu" = Leucin, ...). The three codons TAA, TAG, and TGA are special: instead of coding for an amino acid, they are used to indicate that the protein ends.

In order to make protein, DNA is first copied into a similar molecule called messenger RNA (abbreviated mRNA) in a process called *transcription*. It is the RNA that is *translated* into protein. The entire process is referred to as *expression*. Proteins can be structural elements or perform complex tasks (such as regulation of expression) by interacting with the many molecules and complexes in cells. Thus, the genome is a blueprint for life. An understanding of the genes, the function of their proteins, and their expression patterns is fundamental to biology.

The human genome contains approximately 25,000 genes, although the exact number has still not been determined. While there are experimental methods for

validating and discovering genes, there is still no known high-throughput technology for accurately identifying all the genes in a genome. The computational problem of identifying genes, the *gene finding problem*, is an active area of research. One of the main difficulties lies in the fact that only a small portion of any genome is genic. For instance, less than 5% of the human genome is known to be functional. In section 4 we discuss this problem and the role of probabilistic models in formulating statistically sound methods for distinguishing genes from nongenic sequence. The models of choice, HMMs, allow for the integration of diverse biological information (such as the genetic code and the structure of genes) and yet are suitable for designing efficient algorithms. By virtue of being algebraic varieties, they provide a key example of the link connecting algebra, statistics, and genomics. Nevertheless, the current understanding of genes is not sufficient to allow for the *ab-initio* identification of all the genes in a genome, and it is through comparison with other genomes that the genes are revealed [3].

The differences between the genomes of individuals in a population are small and are primarily due to recombination events (part of the process by which two copies of parental chromosomes are merged in the offspring). On the other hand, the genomes of different *species* (classes of organisms that can produce offspring together) tend to be much more divergent. Genome differences between species can be explained by many biological events including:

- *Genome rearrangement*—comparing chromosomes of related species reveals large segments that have been reversed and flipped (*inversions*), segments that have been moved (*transpositions*), *fusions* of chromosomes, and other large scale events. The underlying biological mechanisms are poorly understood [45, 49].
- *Duplications and loss*—some genomes have undergone whole genome duplications. This process was recently demonstrated for yeast [36]. Individual chromosomes or genes may also be duplicated. Duplication events are often accompanied by *gene loss*, as redundant genes slowly lose or adapt their function over time [23].
- *Parasitic expansion*—large sections of genomes are repetitive, consisting of elements which can duplicate and reintegrate into a genome.
- *Point mutation, insertion, and deletion*—DNA sequences mutate, and in nonfunctional regions these mutations accumulate over time. Such regions are also likely to exhibit deletions; for example, strand slippage during replication can lead to an incorrect copy number for repeated bases.

Accurate mathematical models for sequence alignment and evolution, our topics in sections 5–7, have to take these processes into consideration.

Two distinct DNA bases that share a common ancestor are called *homologous*. Homologous bases can be related via speciation and duplication events, and are therefore divided into two classes: *orthologous* and *paralogous*. Orthologous bases are descendant from a single base in an ancestral genome that underwent a speciation event, whereas two paralogous bases correspond to two distinct bases in a single ancestral genome that are related via a duplication. Because we cannot sequence ancestral genomes, it is never possible to formally prove that two DNA bases are homologous. However, statistical arguments can show that it is extremely likely that two bases are homologous, or even orthologous. The problem of identifying homologous bases between genomes of related species is known as the *alignment problem*. We shall discuss this in section 5.

The alignment of genomes is the first step in identifying highly conserved sequences that point to the small fraction of the genome that is under selection, and therefore likely to be functional. Although the problem of sequence alignment is mathematically and computationally challenging, proposed homologous sequences can be rapidly and independently validated (it is easy to check whether two sequences align once they have been identified), and the regions can often be tested in a molecular biology laboratory to determine their function. In other words, sequence alignment reveals concrete verifiable evidence for evolutionary selection and often results in testable hypotheses.

As a focal point for our discussion, we present a specific DNA sequence of length 42. This sequence was found in the fall of 2003 as a byproduct of computational work conducted by Lior Pachter's group at Berkeley [10]. Whole genome alignments were found and analyzed for human (hs), chimpanzee (pt), mouse (mm), rat (rn), dog (cf), chicken (gg), frog (xt), zebra-fish (dr), fugu-fish (tr), and tetraodon (tn) genomes. The abbreviations refer to the Latin names of these organisms. They will be used in Table 8.1 and Figure 8.1. From alignments of the ten genomes, the following hypothesis was derived, which we state in the form of a mathematical conjecture.

CONJECTURE 1 (the "Meaning of Life"). *The sequence of* 42 *bases*

(2.1) TTTAATTGAAAGAAGTTAATTGAATGAAAATGATCAACTAAG

was present in the genome of the ancestor of all vertebrates, and it has been completely conserved to the present time (i.e., none of the bases have been mutated, nor have there been any insertions or deletions).

The identification of such a sequence requires a highly nontrivial computation: the alignment of ten genomes (including mammalian genomes close to 3 billion bases in length) and subsequent analysis to identify conserved orthologous regions within the alignment [63]. Using the tools described in section 8, one checks that the sequence (2.1) is present in all ten genomes. For instance, in the human genome (May 2004 version), the sequence occurs on chromosome 7 in positions 156501197–156501238. By examining the alignment, one verifies that, with very high probability, the regions containing this sequence in all ten genomes are orthologous. Furthermore, the implied claim that (2.1) occurs in all present-day vertebrates can, in principle, be tested.

Identifying and analyzing sequences such as (2.1) is important because they are highly conserved yet often nongenic [7]. One of the ongoing mysteries in biology is to unravel the function of the parts of the genome that are nongenic and yet very conserved. The extent of conservation points to the possibility of critical functions within the genome. Recent studies have pointed to the association of highly conserved elements with developmental genes [48, 62].

In 2003, the sequence (2.1) appeared to be the longest completely conserved sequence among the vertebrates. We were amused to find that its length was 42. In light of [1], it was decided to name this DNA sequence "The Meaning of Life." It may be a coincidence that the segment above contains two copies of the *motif* `TTAATTGAA`, but this motif may also have some function (for example, it may be bound by a protein). Indeed, the identification of such elements is the first step toward understanding the complex regulatory code of the genome.

The conjecture was formulated in the spring of 2004 and it was circulated in the first `arXiv` version of this paper. In the fall of 2004, Drton, Eriksson, and Leung [21] conducted a new study based on improved alignments. Their work, and similar studies by other groups [51], have now led to the identification of longer sequences

with similar properties. Thus, the Meaning of Life sequence no longer holds the record in terms of length. However, since Conjecture 1 has been inspiration for our group, and it still remains open today, we decided to stick with this example. It needs to be emphasized that disproving Conjecture 1 would not invalidate any of the methodology presented in this article. For a biological perspective we refer to [21].

3. Codons. Because of the genetic code, the set Ω^3 of all three-letter words over the alphabet $\Omega = \{A, C, G, T\}$ plays a special role in molecular biology. As was discussed in section 2, these words are called codons, with each triplet coding for one of 20 amino acids (Table 2.1). The map from 64 codons to 20 amino acids is not injective, and so multiple codons code for the same amino acid. Such codons are called *synonymous*. Eight amino acids have the property that the synonymous codons that code for them all agree in the first two positions. The third positions of such codons are called *four-fold degenerate*. The translation of a series of codons in a gene (typically a few hundred) results in a three-dimensional folded protein.

A *model for codons* is a statistical model whose state space is the 64-element set Ω^3. Selecting a model means specifying a family of probability distributions $p = (p_{IJK})$ on Ω^3. Each probability distribution p is a $4 \times 4 \times 4$-table of nonnegative real numbers which sum to one. Geometrically, a distribution on codons is a point p in the 63-dimensional probability simplex

$$\Delta_{63} = \left\{ p \in \mathbb{R}^{\Omega^3} \ : \ \sum_{IJK \in \Omega^3} p_{IJK} = 1 \ \text{ and } \ p_{IJK} \geq 0 \ \text{ for all } \ IJK \in \Omega^3 \right\}.$$

A model for codons is hence nothing but a subset $\mathcal{M}$ of the simplex Δ_{63}. Statistically meaningful models are usually given in parametric form. If the number of parameters is d, then there is a set $\mathcal{P} \subset \mathbb{R}^d$ of allowed parameters, and the model $\mathcal{M}$ is the image of a map ϕ from $\mathcal{P}$ into Δ_{63}. We illustrate this statistical point of view by means of a very simple independence model.

Models for codons have played a prominent role in the work of Samuel Karlin, who was one of the mathematical pioneers in this field. One instance of this is the *genome signature* in [13]. We refer to [44, Example 4.3] for a discussion of this model and more recent work on codon usage in genomes.

Consider a DNA sequence of length $3m$ which has been grouped into m consecutive codons. Let u_{IJK} denote the number of occurrences of a particular codon IJK. Then our data are the $4 \times 4 \times 4$-table $u = (u_{IJK})$. The entries of this table are nonnegative integers, and if we divide each entry by m, we then get a new table $\frac{1}{m} \cdot u$ which is a point in the probability simplex Δ_{63}. This table is the *empirical distribution of codons* in the given sequence.

Let $\mathcal{M}$ be the statistical model which stipulates that, for the sequence under consideration, the first two positions in a codon are independent of the third position. We may wish to test whether this independence model fits our data u. This question makes sense in molecular biology because many of the amino acids are uniquely specified by the first two positions in any codon which represents that particular amino acid (see Table 2.1). Therefore, third positions in synonymous codons tend to be independent of the first two.

Our independence model $\mathcal{M}$ has 18 free parameters. The set of allowed parameters is an 18-dimensional convex polytope, namely, it is the product

$$\mathcal{P} = \Delta_{15} \times \Delta_3.$$

Here Δ_{15} is the 15-dimensional simplex consisting of probability distributions $\alpha = (\alpha_{IJ})$ on Ω^2, and Δ_3 is the tetrahedron consisting of probability distributions $\beta = (\beta_K)$ on Ω. Our model $\mathcal{M}$ is parameterized by the map

$$\phi \,:\, \mathcal{P} \,\rightarrow\, \Delta_{63}\,, \quad \phi((\alpha,\beta))_{IJK} = \alpha_{IJ} \cdot \beta_K.$$

Hence $\mathcal{M} = \text{image}(\phi)$ is an 18-dimensional algebraic subset inside the 63-dimensional simplex. To test whether a given $4 \times 4 \times 4$-table p lies in $\mathcal{M}$, we write that table as a two-dimensional matrix with 16 rows and 4 columns:

$$p' = \begin{pmatrix} p_{AAA} & p_{AAC} & p_{AAG} & p_{AAT} \\ p_{ACA} & p_{ACC} & p_{ACG} & p_{ACT} \\ p_{AGA} & p_{AGC} & p_{AGG} & p_{AGT} \\ p_{ATA} & p_{ATC} & p_{ATG} & p_{ATT} \\ p_{CAA} & p_{CAC} & p_{CAG} & p_{CAT} \\ \vdots & \vdots & \vdots & \vdots \\ p_{TTA} & p_{TTC} & p_{TTG} & p_{TTT} \end{pmatrix}.$$

Linear algebra furnishes the following characterizations of our model.

PROPOSITION 2. *For a point $p \in \Delta_{63}$, the following conditions are equivalent:*

1. *The distribution p lies in the model $\mathcal{M}$.*
2. *The 16×4 matrix p' has rank one.*
3. *All 2×2-minors of the matrix p' are zero.*
4. *$p_{IJK} \cdot p_{LMN} = p_{IJN} \cdot p_{LMK}$ for all nucleotides I, J, K, L, M, N.*

In the language of algebraic geometry, the model $\mathcal{M}$ is known as the *Segre variety*. More precisely, $\mathcal{M}$ is the set of nonnegative real points on the Segre embedding of $\mathbb{P}^{15} \times \mathbb{P}^3$ in $\mathbb{P}^{63}$. Here and throughout, the symbol $\mathbb{P}^m$ denotes the complex projective space of dimension m. One of the points argued in this paper is that many of the more advanced statistical models, such as graphical models [44, section 1.5], actually used in practice by computational biologists are also algebraic varieties with a special combinatorial structure.

Returning to our original biological motivation, we are faced with the following statistics problem. The DNA sequence under consideration is summarized in the data u, and we wish to test whether or not the model $\mathcal{M}$ fits the data. The geometric idea of such a test is to determine whether or not the empirical distribution $\frac{1}{m} \cdot u$ lies close to the Segre variety $\mathcal{M}$. Statisticians have devised a wide range of such tests, each representing a statistically meaningful notion of "proximity to $\mathcal{M}$." These include the χ^2-test, the G^2-test, Fisher's exact test, and others, as explained in standard statistics texts such as [8] or [28]. A useful tool of numerical linear algebra for measuring the distance of a point to the Segre variety is the *singular value decomposition* of the matrix p'. Indeed, p' lies on $\mathcal{M}$ if and only if the second singular value of p' is zero. Singular values provide a good notion of distance between a given matrix and various determinantal varieties such as $\mathcal{M}$.

One key ingredient in statistical tests is *maximum likelihood estimation*. The basic idea is to find those model parameters α_{IJ} and β_K which would best explain the observed data. If we consider all possible genome sequences of length $3m$, then the likelihood of observing our particular data u equals

$$\gamma \cdot \prod_{IJK \in \Omega^3} p_{IJK}^{u_{IJK}},$$

where γ is a combinatorial constant. This expression is a function of (α, β), called the *likelihood function*. We wish to find the point in our parameter domain $\mathcal{P} = \Delta_{15} \times \Delta_3$

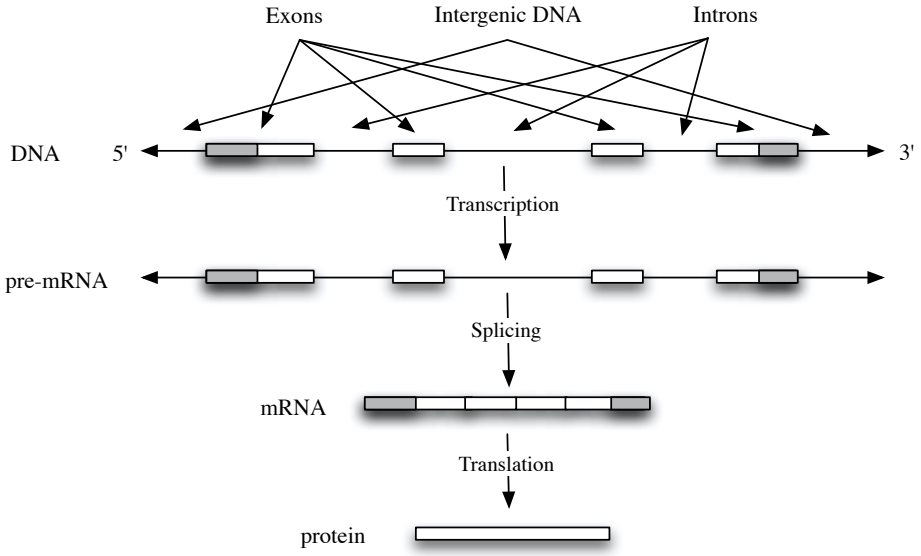

Fig. 4.1 *Structure of a gene.*

which maximizes this function. The solution $(\hat{\alpha}, \hat{\beta})$ to this nonlinear optimization problem is said to be the *maximum likelihood estimate* for the data u. In our independence model, the likelihood function is convex, and it is easy to write down the global maximum explicitly:

$$\hat{\alpha}_{IJ} = \frac{1}{m} \sum_{K \in \Omega} u_{IJK} \qquad \text{and} \qquad \hat{\beta}_K =, \frac{1}{m} \sum_{IJ \in \Omega^2} u_{IJK}.$$

In general, the likelihood function of a statistical model will not be convex, and there is no easy formula for writing the maximum likelihood estimate as a function of the data. In practice, numerical hill-climbing methods are used to solve this optimization problem, but, of course, there is no guarantee that a local maximum found by such methods is actually the global maximum.

4. Gene Finding. In order to find genes in DNA sequences, it is necessary to identify structural features and sequence characteristics that distinguish genic sequence from nongenic sequence. We begin by describing more of the detail of gene structure which is essential in developing probabilistic models.

Genes are not contiguous subsequences of the genome, but rather are split into pieces called *introns* and *exons*. After transcription, introns are spliced out and only the remaining exons are used in translation (Figure 4.1). Not all of the sequence in the exons is translated; the initial and terminal exons may consist of *untranslated regions* (indicated in gray in the figure). Since the genetic code is in (nonoverlapping) triplets, it follows that the lengths of the translated portions of the exons must sum to 0 mod 3. In addition to the exon-intron structure of genes, there are known sequence signals. The codon ATG initiates translation, and thus is the first codon following the untranslated portion of the initial exons. The final codon in a gene must be one of TAG, TAA, or TGA, as indicted in Table 2.1. These codons signal the translation

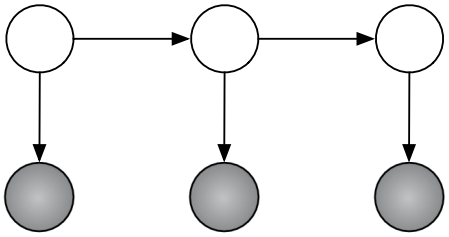

Fig. 4.2 *The HMM of length three.*

machinery to stop. There are also sequence signals at the intron-exon boundaries: GT at the $5'$ end of an intron and AG at the $3'$ end.

A *hidden Markov model* (HMM) is a probabilistic model that allows for simultaneous modeling of the bases in a DNA sequence of length n and the structural features associated with that sequence. The HMM consists of n *observed* random variables $Y_1, \ldots, Y_n$ taking on l possible states, and n *hidden* random variables $X_1, \ldots, X_n$ taking on k possible states. In the context of phylogenomics, the observed variables Y_i usually have $l = 4$ states, namely, $\Omega = \{A, C, G, T\}$. The hidden random variables X_i serve to model features associated with the sequence which is generated by $Y_1, Y_2, \ldots, Y_n$. An oversimplified scenario is $k = 2$, with the set of hidden states being $\Theta = \{\, exon,\, intron\}$.

The characteristic property of an HMM is that the distributions of the Y_i depend on the X_i, while the X_i form a *Markov chain*. This is illustrated for $n = 3$ in Figure 4.2, where the unshaded circles represent the hidden variables X_1, X_2, X_3 and the shaded circles represent the observed variables Y_1, Y_2, Y_3.

Computational biologists use HMMs to annotate DNA sequences. The basic idea is this: it is postulated that the bases are instances of the random variables $Y_1, \ldots, Y_n$, and the problem is to identify the most likely assignments of states to $X_1, \ldots, X_n$ that could be associated with the observations. In gene finding, *homogeneous HMMs* are used. This means that all transition probabilities $X_i \to X_{i+1}$ are given by the same $k \times k$-matrix $S = (s_{ij})$, and all the transitions $X_i \to Y_i$ are given by another $k \times 4$-matrix $T = (t_{ij})$. Here s_{ij} represents the probability of transitioning from hidden state i to hidden state j; for instance, if $k = 2$, then $i, j \in \Theta = \{\, exon,\, intron\}$. The parameter t_{ij} represents the probability that state $i \in \Theta$ outputs letter $j \in \Omega$.

In practice, the parameters s_{ij} and t_{ij} range over real numbers satisfying

$$s_{ij}\,,\; t_{ij} \;\geq\; 0 \quad \text{and} \quad \sum_{j \in \Theta} s_{1j} \;=\; \sum_{j \in \Omega} t_{1j} \;=\; 1. \tag{4.1}$$

However, just like in our discussion of the Segre variety in section 3, we may relax the requirements (4.1) and allow the parameters to be arbitrary complex numbers. This leads to the following algebraic representation [42, section 2].

PROPOSITION 3. *The homogeneous HMM is the image of a map* $\phi : \mathbb{C}^{k(k+l)} \to \mathbb{C}^{l^n}$, *where each coordinate of* ϕ *is a bihomogeneous polynomial of degree* $n-1$ *in the transition probabilities* s_{ij} *and degree* n *in the output probabilities* t_{ij}.

The coordinate ϕ_σ of the map ϕ indexed by a particular DNA sequence $\sigma \in \Omega^n$ represents the probability that the HMM generates the sequence σ. The following

explicit formula for that probability establishes Proposition 3:

$$\phi_\sigma = \sum_{i_1 \in \Theta} t_{i_1\sigma_1} \Big(\sum_{i_2 \in \Theta} s_{i_1 i_2} t_{i_2 \sigma_2} \Big(\sum_{i_3 \in \Theta} s_{i_2 i_3} t_{i_3 \sigma_3} \Big(\sum_{i_4 \in \Theta} s_{i_3 i_4} t_{i_4 \sigma_4} \, (\, \cdots \,) \Big) \Big) \Big). \tag{4.2}$$

The expansion of this polynomial has k^n terms:

$$t_{i_1\sigma_1} s_{i_1 i_2} t_{i_2\sigma_2} s_{i_2 i_3} t_{i_3\sigma_3} \cdots s_{i_{n-1} i_n} t_{i_n \sigma_n}. \tag{4.3}$$

For any fixed parameters one wishes to determine a string $\widehat{\mathbf{i}} = (i_1, i_2, \ldots, i_n) \in \Theta^n$ which indexes a term (4.3) of largest numerical value among all k^n terms of ϕ_σ. (If there is more than one string with maximum value, then we break ties lexicographically.) We call $\widehat{\mathbf{i}}$ the *explanation* of the observation σ. In our example ($k = 2, l = 4$), the explanation $\widehat{\mathbf{i}}$ of a DNA sequence σ is an element of $\Theta^n = \{ \mathit{exon}, \mathit{intron} \}^n$. It reveals the crucial information of Figure 4.1, namely, the location of the exons and introns. In summary, the DNA sequence to be annotated by an HMM corresponds to the observation $\sigma \in \Omega^n$, and the explanation $\widehat{\mathbf{i}}$ is the gene prediction. Thus *gene finding* means nothing but computing the output $\widehat{\mathbf{i}}$ from the input σ.

In real-world applications, the integer n may be quite large. It is not uncommon to annotate DNA sequences of length $n \geq$ 1,000,000. The size k^n of the search space for finding the explanation is enormous (exponential in n). Fortunately, the recursive decomposition in (4.2), reminiscent of *Horner's Rule*, allows us to evaluate a multivariate polynomial with exponentially many terms in linear time (in n). In other words, for given numerical parameters s_{ij} and t_{ij}, we can compute the probability $\phi_\sigma(s_{ij}, t_{ij})$ quite efficiently.

Similarly, the explanation $\widehat{\mathbf{i}}$ of an observed DNA sequence σ can be computed in linear time. This is done using the *Viterbi algorithm*, which evaluates

$$\max_{i_1 \in \Theta} T_{i_1\sigma_1} + \Big(\max_{i_2 \in \Theta} S_{i_1 i_2} + T_{i_2 \sigma_2} + \Big(\max_{i_3 \in \Theta} S_{i_2 i_3} + T_{i_3 \sigma_3} + \Big(\max_{i_4 \in \Theta} S_{i_3 i_4} + T_{i_4 \sigma_4} + \, (\cdots) \Big) \Big) \Big),$$

where $S_{ij} = \log(s_{ij})$ and $T_{ij} = \log(t_{ij})$. This expression is a piecewise linear convex function on $\mathbb{R}^{k(k+l)}$, known as the *tropicalization* of the polynomial ϕ_σ. Indeed, evaluating this expression requires exactly the same operations as evaluating ϕ_σ, with the only difference that we are replacing ordinary arithmetic by the *tropical semiring*. The tropical semiring (also known as the *max-plus* algebra) consists of the real numbers $\mathbb{R}$ together with an extra element ∞, where the arithmetic operations of addition and multiplication are redefined to be *max* (or equivalently *min*) and *plus*, respectively. The tropical semiring and its use in dynamic programming optimizations are explained in [44, section 2.1].

Every choice of parameters (s_{ij}, t_{ij}) specifies a *gene finding function*

$$\Omega^n \to \Theta^n, \quad \sigma \mapsto \widehat{\mathbf{i}},$$

which takes a sequence σ to its explanation $\widehat{\mathbf{i}}$. The number of all functions from Ω^n to Θ^n equals $2^{n \cdot 4^n}$ and hence grows double-exponentially in n. However, the vast majority of these functions are not gene finding functions. The following remarkable complexity result was proved by Elizalde [24].

THEOREM 4. *The number of gene finding functions grows at most polynomially in the sequence length n.*

As an illustration consider the $n = 3$ example visualized in Figure 4.2. There are $8^{64} = 6.277 \cdot 10^{57}$ functions $\{A, C, G, T\}^3 \to \{exon, intron\}^3$ but only a tiny fraction of these are gene finding functions. (It would be interesting to determine the exact number.) It is an open problem to give a combinatorial characterization of gene finding functions and to come up with accurate lower and upper bounds for their number as n grows.

For gene finding HMMs, it is always the case that l is small and fixed (usually, $l = 4$), and n is large. However, the size of k or structure of the state space for the hidden variables X_i tends to vary a lot. While the $k = 2$ used in our discussion of gene finding functions was meant to be just an illustration, a biologically meaningful gene finding model could work with just three hidden states: one for introns, one for exons, and one for intergenic sequences. However, in order to enforce the constraint that the sum of the lengths of the exons is 0 mod 3, a more complicated hidden state space is necessary. Solutions to this problem were given in [12, 37].

We conclude this section with a brief discussion of the important problem of *estimating parameters* for HMMs. Indeed, so far nothing has been said about how the values of the parameters s_{ij} and t_{ij} are to be chosen when running the Viterbi algorithm. Typically, this choice involves a combination of biological and statistical considerations. Let us concentrate on the latter aspect.

Recall that maximum likelihood estimation is concerned with finding parameters for a statistical model which best explain the observed data. As was the case for the codon model (section 3), the maximum likelihood estimate is an algebraic function of the data. In contrast to what we did at the end of section 3, it is now prohibitive to locate the global maximum in the polytope (4.1). The *expectation-maximization* (EM) algorithm is a general technique used by statisticians to find local maxima of the likelihood function [44, section 1.3]. For HMMs, this algorithm is also known as the *Baum–Welch algorithm*. It takes advantage of the recursive decomposition in (4.2) and it is fast (linear in n). The widely used book [18] provides a good introduction to the use of the Baum–Welch algorithm in training HMMs for biological sequence applications. The connection between the EM algorithm and the Baum–Welch algorithm is explained in detail in [30]. In order to understand the performance of EM or to develop more global methods [14], it would be desirable to obtain upper and lower bounds on the algebraic degree [33] of the maximum likelihood estimate.

5. Sequence Alignment. Although tools such as the HMM are important for modeling and analyzing individual genome sequences, the essence of phylogenomics lies in the power of *sequence comparison*. Because functional sequences tend to accumulate fewer mutations over time, it is possible, by comparing genomes, to identify and characterize such sequences much more effectively.

In this section we examine models for sequence evolution that allow for insertions, deletions, and mutations in the special case of two genomes. These are known as pairwise sequence alignment models. The specific model to be discussed here is the *pair HMM*. In the subsequent section we shall examine phylogenetic models for more than two DNA sequences.

We have already seen two instances of statistical models that are represented by polynomials in the model parameters (the codon model and the HMM). Models for pairwise sequence alignment are also specified by polynomials, and are in fact close relatives of HMMs. What distinguishes the sequence alignment problem is an extra layer of complexity which arises from a combinatorial explosion in the number of possible alignments between sequences. Here we describe one of the simplest alignment

Table 5.1 *Alignments for a pair of sequences of length* 2 *and* 3.

IIIDD	$(\cdots ij\,,\, klm\cdot\cdot)$	$t_{Ik}s_{II}t_{Il}s_{II}t_{Im}s_{ID}t_{Di}s_{DD}t_{Dj}$
IIDID	$(\cdots i\cdot j\,,\, kl\cdot m\cdot)$	$t_{Ik}s_{II}t_{Il}s_{ID}t_{Di}s_{DI}t_{Im}s_{ID}t_{Dj}$
IIDDI	$(\cdots ij\cdot\,,\, kl\cdot\cdot m)$	$t_{Ik}s_{II}t_{Il}s_{ID}t_{Di}s_{DD}t_{Dj}s_{DI}t_{Im}$
IDIID	$(\cdot i\cdot\cdot j\,,\, k\cdot lm\cdot)$	$t_{Ik}s_{ID}t_{Di}s_{DI}t_{Il}s_{II}t_{Im}s_{ID}t_{Dj}$
IDIDI	$(\cdot i\cdot j\cdot\,,\, k\cdot l\cdot m)$	$t_{Ik}s_{ID}t_{Di}s_{DI}t_{Il}s_{ID}t_{Dj}s_{DI}t_{Im}$
IDDII	$(\cdot ij\cdot\cdot\,,\, k\cdot\cdot lm)$	$t_{Ik}s_{ID}t_{Di}s_{DD}t_{Dj}s_{DI}t_{Il}s_{II}t_{Im}$
DIIID	$(i\cdot\cdot\cdot j\,,\, \cdot klm\cdot)$	$t_{Di}s_{DI}t_{Ik}s_{II}t_{Il}s_{II}t_{Im}s_{ID}t_{Dj}$
DIIDI	$(i\cdot\cdot j\cdot\,,\, \cdot kl\cdot m)$	$t_{Di}s_{DI}t_{Ik}s_{II}t_{Il}s_{ID}t_{Dj}s_{DI}t_{Im}$
DIDII	$(i\cdot j\cdot\cdot\,,\, \cdot k\cdot lm)$	$t_{Di}s_{DI}t_{Ik}s_{ID}t_{Dj}s_{DI}t_{Il}s_{II}t_{Im}$
DDIII	$(ij\cdot\cdot\cdot\,,\, \cdot\cdot klm)$	$t_{Di}s_{DD}t_{Dj}s_{DI}t_{Ik}s_{II}t_{Il}s_{II}t_{Im}$
MIID	$(i\cdot\cdot j\,,\, klm\cdot)$	$t_{Mik}s_{MI}t_{Il}s_{II}t_{Im}s_{ID}t_{Dj}$
MIDI	$(i\cdot j\cdot\,,\, kl\cdot m)$	$t_{Mik}s_{MI}t_{Il}s_{ID}t_{Dj}s_{DI}t_{Im}$
MDII	$(ij\cdot\cdot\,,\, k\cdot lm)$	$t_{Mik}s_{MD}t_{Dj}s_{DI}t_{Il}s_{II}t_{Im}$
IMID	$(\cdot i\cdot j\,,\, klm\cdot)$	$t_{Ik}s_{IM}t_{Mil}s_{MI}t_{Im}s_{ID}t_{Dj}$
IMDI	$(\cdot ij\cdot\,,\, kl\cdot m)$	$t_{Ik}s_{IM}t_{Mil}s_{MD}t_{Dj}s_{DI}t_{Im}$
IIMD	$(\cdot\cdot ij\,,\, klm\cdot)$	$t_{Ik}s_{II}t_{Il}s_{IM}t_{Mim}s_{MD}t_{Dj}$
IIDM	$(\cdot\cdot ij\,,\, kl\cdot m)$	$t_{Ik}s_{II}t_{Il}s_{ID}t_{Di}s_{DM}t_{Mjm}$
IDMI	$(\cdot ij\cdot\,,\, k\cdot lm)$	$t_{Ik}s_{ID}t_{Di}s_{DM}t_{Mjl}s_{MI}t_{Im}$
IDIM	$(\cdot i\cdot j\,,\, k\cdot lm)$	$t_{Ik}s_{ID}t_{Di}s_{DI}t_{Il}s_{IM}t_{Mjm}$
DMII	$(ij\cdot\cdot\,,\, \cdot klm)$	$t_{Di}s_{DM}t_{Mjk}s_{MI}t_{Il}s_{II}t_{Im}$
DIMI	$(i\cdot j\cdot\,,\, \cdot klm)$	$t_{Di}s_{DI}t_{Ik}s_{IM}t_{Mjl}s_{MI}t_{Im}$
DIIM	$(i\cdot\cdot j\,,\, \cdot klm)$	$t_{Di}s_{DI}t_{Ik}s_{II}t_{Il}s_{IM}t_{Mjm}$
MMI	$(ij\cdot\,,\ klm)$	$t_{Mik}s_{MM}t_{Mjl}s_{MI}t_{Im}$
MIM	$(i\cdot j\ ,\ klm)$	$t_{Mik}s_{MI}t_{Il}s_{IM}t_{Mjm}$
IMM	$(\cdot ij\ ,\ klm)$	$t_{Ik}s_{IM}t_{Mil}s_{MM}t_{Mjm}$

models (for a pair of sequences), with a view toward connections with tree models and algebraic statistics.

Given two sequences $\sigma^1 = \sigma^1_1\sigma^1_2\cdots\sigma^1_n$ and $\sigma^2 = \sigma^2_1\sigma^2_2\cdots\sigma^2_m$ over the alphabet $\Omega = \{A, C, G, T\}$, an *alignment* is a string over the auxiliary alphabet $\{M, I, D\}$ such that $\#M + \#D = n$ and $\#M + \#I = m$. Here $\#M, \#I, \#D$ denote the number of characters M, I, D in the word, respectively. An alignment records the "edit steps" from the sequence σ^1 to the sequence σ^2, where edit operations consist of changing characters, preserving them, or inserting/deleting them. An I in the alignment string corresponds to an insertion from the first sequence to the second, a D is a deletion from the first sequence to the second, and an M is either a character change or lack thereof. The set $\mathcal{A}_{n.m}$ of all alignments depends only on the integers n and m, and not on σ^1 and σ^2.

PROPOSITION 5. *The cardinality of the set $\mathcal{A}_{n.m}$ of all alignments can be computed as the coefficient of the monomial $x^m y^n$ in the generating function*

$$\frac{1}{1-x-y-xy} = 1 + x + y + x^2 + 3xy + y^2 + \cdots + x^5 + 9x^4y + 25x^3y^2 + \cdots.$$

These cardinalities $|\mathcal{A}_{n,m}|$ are known as *Delannoy numbers* in combinatorics [55, section 6.3]. For instance, there are $|\mathcal{A}_{2,3}| = 25$ alignments of two sequences of length two and three. They are listed in Table 5.1 below.

The *pair HMM* is visualized graphically in Figure 5.1. The hidden random variables (unshaded nodes forming the Markov chain) take on the values M, I, D. Depending on the state at a hidden node, either one or two characters are generated; in this way, pair HMMs differ from standard HMMs. The squares around the observed states (called plates) are used to indicate that the number of characters generated

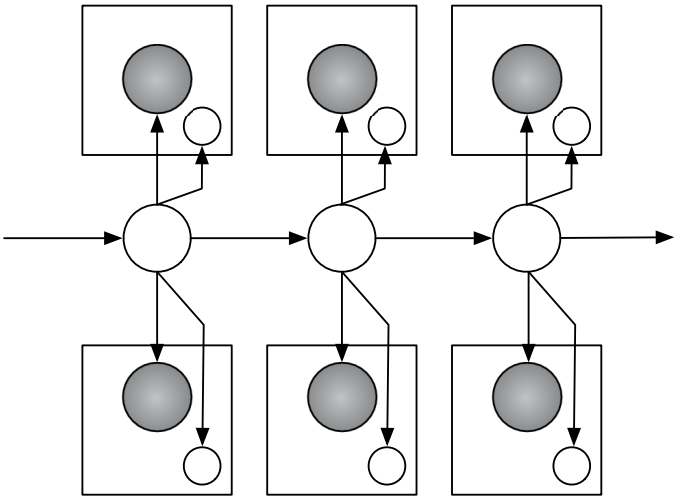

Fig. 5.1 *A pair HMM for sequence alignment.*

may vary depending on the hidden state. The number of characters generated is a random variable, indicated by unshaded nodes within the plates (called class nodes). In pair HMMs, the class nodes take on the values 0 or 1 corresponding to whether or not a character is generated. Pair HMMs are therefore HMMs where the structure of the model depends on the assignments to the hidden states. The graphical model structure of pair HMMs is explained in more detail in [2].

The next proposition gives the algebraic representation of the pair HMM. For a given alignment $\mathbf{a} \in \mathcal{A}_{n,m}$, we denote the jth character in $\mathbf{a}$ by a_j, we write $a[i]$ for $\#M + \#D$ in the prefix $a_1 a_2 \ldots a_i$, and we write $a\langle j \rangle$ for $\#M + \#I$ in the prefix $a_1 a_2 \ldots a_j$. Let σ^1 and σ^2 be two DNA sequences of lengths n, m, respectively. Then the probability that our model generates these two sequences equals

$$\phi_{\sigma^1,\sigma^2} = \sum_{\mathbf{a}\in\mathcal{A}_{n,m}} t_{a_1}(\sigma^1_{a[1]}, \sigma^2_{a\langle 1\rangle}) \cdot \prod_{i=2}^{|\mathbf{a}|} s_{a_{i-1}a_i} \cdot t_{a_i}(\sigma^1_{a[i]}, \sigma^2_{a\langle i\rangle}), \tag{5.1}$$

where the parameter $s_{a_{i-1}a_i}$ is the transition probability from state a_{i-1} to a_i, and the parameter $t_{a_i}(\sigma^1_{a[i]}, \sigma^2_{a\langle i\rangle})$ is the output probability for a given state a_i and the indicated output characters on the strings σ^1 and σ^2.

PROPOSITION 6. *The pair HMM for sequence alignment is the image of a polynomial map* $\phi : \mathbb{C}^{33} \to \mathbb{C}^{4^{n+m}}$. *The coordinates of the map* ϕ *are the polynomials of degree* $\leq 2n + 2m - 1$ *which are given in* (5.1).

We need to explain why the number of parameters in our representation of the pair HMM is 33. First, there are nine parameters

$$S = \begin{pmatrix} s_{MM} & s_{MI} & s_{MD} \\ s_{IM} & s_{II} & s_{ID} \\ s_{DM} & s_{DI} & s_{DD} \end{pmatrix}$$

which play the same role as in section 4, namely, they represent transition probabilities in the Markov chain. There are 16 parameters $t_M(a, b) =: t_{Mab}$ for the probability

that letter a in σ^1 is matched with letter b in σ^2. The insertion parameters $t_I(a,b)$ depend only on the letter b, and the deletion parameters $t_D(a,b)$ depend only on the letter a, so there are only 8 of these parameters. Hence the total number of (complex) parameters is $9+16+8 = 33$. Of course, in our applications, probabilities are nonnegative reals that sum to one, so we get a reduction in the number of parameters, just like in (4.1). In the upcoming example, which explains the algebraic representation of Proposition 6, we use the abbreviations t_{Ib} and t_{Da} for these parameters.

Consider two sequences $\sigma^1 = ij$ and $\sigma^2 = klm$ of length $n = 2$ and $m = 3$ over the alphabet $\Omega = \{A, C, G, T\}$. The number of alignments is $|\mathcal{A}_{2,3}| = 25$, and they are listed in Table 5.1. For instance, the alignment $MIID$, here written $(\, i \cdot\cdot j \,,\, klm \cdot\,)$, corresponds to $\begin{smallmatrix} \mathtt{i} - - \mathtt{j} \\ \mathtt{k\,l\,m} - \end{smallmatrix}$ in standard genomics notation.

The polynomial ϕ_{σ^1,σ^2} is the sum of the 25 monomials (of degree $9, 7, 5$) in the rightmost column. Thus the pair HMM presented in Table 5.1 is nothing but a polynomial map

$$\phi \,:\, \mathbb{C}^{33} \,\rightarrow\, \mathbb{C}^{1024}.$$

Statistics is all about making inferences. We shall now explain how this is done with this model. For any fixed parameters $s_{\cdot\cdot}$ and $t_{\cdot\cdot}$, one wishes to determine the alignment $\widehat{\mathbf{a}} \in \mathcal{A}_{n,m}$ which indexes the term of largest numerical value among the many terms (see Proposition 5) of the polynomial ϕ_{σ_1,σ_2}. (If there is more than one alignment with maximum value, then we break ties lexicographically.) We call $\widehat{\mathbf{a}}$ the *explanation* of the observation (σ^1, σ^2).

The explanation for a pair of DNA sequences can be computed in polynomial time (in their lengths n and m) using a variant of the Viterbi algorithm. Just like in the previous section, the key idea is to *tropicalize* the coordinate polynomials (5.1) of the statistical model in question. Namely, we compute

$$\max_{\mathbf{a}\in\mathcal{A}_{n,m}} \; T_{a_1}(\sigma^1_{a[1]}, \sigma^2_{a\langle 1\rangle}) + \sum_{i=2}^{|\mathbf{a}|} S_{a_{i-1}a_i} + T_{a_i}(\sigma^1_{a[i]}, \sigma^2_{a\langle i\rangle}), \tag{5.2}$$

where $S_{\cdot\cdot} = \log(s_{\cdot\cdot})$ and $T_{\cdot\cdot} = \log(t_{\cdot\cdot})$. The "arg max" of this piecewise linear convex function is the optimal alignment $\widehat{\mathbf{a}}$. Inference in the pair HMM means computing the optimal alignment of two observed DNA sequences. In other words, by inference we mean evaluating the *alignment function*

$$\Omega^n \times \Omega^m \;\rightarrow\; \mathcal{A}_{n,m}\,, \quad (\sigma^1, \sigma^2) \mapsto \widehat{\mathbf{a}}.$$

There are doubly-exponentially many functions from $\Omega^n \times \Omega^m$ to $\mathcal{A}_{n,m}$, but, by Elizalde's *few inference functions theorem* [24], at most polynomially many of them are alignment functions. Like for gene finding functions (cf. Theorem 4), it is an open problem to characterize alignment functions.

The function $\mathbb{R}^{33} \rightarrow \mathbb{R}$ given in (5.2) is the support function of a convex polytope in $\mathbb{R}^{33}$, namely, the *Newton polytope* of the polynomial ϕ_{σ^1,σ^2}. The vertices of this polytope correspond to all optimal alignments of the sequences σ^1, σ^2 with respect to all possible choices of the parameters, and the normal fan of the polytope divides the logarithmic parameter space into regions which yield the same optimal alignment. This can be used for analyzing the sensitivity of alignments to parameters, and for the computation of posterior probabilities of optimal alignments. The process of

computing this polytope is called *parametric alignment* or *parametric inference*. It is known [27, 43, 60] that parametric inference can be done in polynomial time (in m and n).

An important remark is that the formulation of sequence alignment with pair HMMs is equivalent to combinatorial "scoring schemes" or "generalized edit distances" which can be used to assign weights to alignments [11]. The simplest scoring scheme consists of two parameters: a mismatch score mis, and an indel score gap [29]. The weight of an alignment is the sum of the scores for all positions in the alignment, where a match gets a score of 1. In the case where mis and gap are nonnegative, this is equivalent to specializing the 33 logarithmic parameters $S_{\cdot\cdot} = \log(s_{\cdot\cdot})$ and $T_{\cdot\cdot} = \log(t_{\cdot\cdot})$ of the pair HMM as follows:

$$S_{ij} = 0, \quad T_{Ij} = T_{Di} = -gap \quad \text{for all } i, j,$$

$$T_{Mij} = -1 \quad \text{if } i = j, \qquad \text{and} \qquad T_{Mij} = -mis \quad \text{if } i \neq j.$$

The case where the scoring scheme consists of both positive and negative parameters corresponds to a normalized pair HMM [18]. This specialization of the parameters corresponds to projecting the Newton polytope of ϕ_{σ^1,σ^2} into two dimensions. Parametric alignment means computing the resulting two-dimensional polygon. For two sequences of length n, an upper bound on the number of vertices in the polygon is $O(n^{2/3})$. We have observed that for biological sequences the number may be much smaller. See [27] for a survey from the perspective of computational geometry.

In the strict technical sense, our polynomial formulation (5.1) is not needed to derive or analyze combinatorial algorithms for sequence alignment. However, the translation from algebraic geometry (5.1) to discrete optimization (5.2) offers much more than just esthetically pleasing formulas. We posit that (tropical) algebraic geometry is a conceptual framework for developing new models and designing new algorithms of practical value for phylogenomics.

6. Models of Evolution. Because organisms from different species cannot produce offspring together, mutations and genome changes that occur within a species are independent of those occurring in another species. There are some exceptions to this statement, such as the known phenomenon of *horizontal transfer* in bacteria which results in the transfer of genetic material between different species; however, we ignore such scenarios in this discussion. We can therefore represent the evolution of species (or phyla) via a tree structure. The study of tree structures in genome evolution is referred to as *phylogenetics*. A phylogenetic X*-tree* is a tree T with all internal vertices of degree at least 3, and with the leaves labeled by a set X which consists of different species. In this section, we assume that T is known and that vertices in T correspond to known speciation events. We begin by describing statistical models of evolution that are used to identify regions between genomes that are under selection.

Evolutionary models attempt to capture three important aspects of evolving sequences: *branch length*, *substitution*, and *mutation*. Consider a single ancestral base b at the root r of a phylogenetic tree T, and assume that there are no insertions or deletions over time. Since the ancestral base changes, it is possible that at two leaves $x, y \in X$ we observe bases $c_1 \neq c_2$. We say that there has been a *substitution* between x and y. In a probabilistic model of evolution, we would like to capture the possibility of change along internal edges of the tree, with the possibility of back substitutions as well. For example, it is possible that $b \to c_1 \to b \to c_1$ along the path from r to x.

DEFINITION 7. *A* rate matrix *(or* Q-matrix*) is a square matrix* $Q = (q_{ij})_{i,j\in\Omega}$ *(with rows and columns indexed by the nucleotides) satisfying the properties*

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

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

$$q_{ii} < 0 \quad \textit{for all} \quad i \in \Omega.$$

Rate matrices capture the notion of *instantaneous rate of mutation*. From a given rate matrix Q one computes the *substitution matrices* $P(t)$ by exponentiation. The entry of $P(t)$ in row b and column c equals the probability that the substitution $b \to \cdots \to c$ occurs in a time interval of length t. We recall the following well-known result about continuous-time Markov models.

PROPOSITION 8. *Let* Q *be any rate matrix and* $P(t) = e^{Qt} = \sum_{i=0}^{\infty} \frac{1}{i!} Q^i t^i$. *Then*

1. $P(s+t) = P(s) + P(t)$;
2. $P(t)$ *is the unique solution to* $P'(t) = P(t) \cdot Q$, $P(0) = \mathbf{1}$ *for* $t \geq 0$;
3. $P(t)$ *is the unique solution to* $P'(t) = Q \cdot P(t)$, $P(0) = \mathbf{1}$ *for* $t \geq 0$.

Furthermore, a matrix Q *is a rate matrix if and only if the matrix* $P(t) = e^{Qt}$ *is a stochastic matrix (nonnegative with row sums equal to one) for every* t.

The simplest model is the *Jukes–Cantor DNA model*, whose rate matrix is

$$Q = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix},$$

where $\alpha \geq 0$ is a parameter. The corresponding substitution matrix equals

$$P(t) = \frac{1}{4} \begin{pmatrix} 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} \end{pmatrix}.$$

The expected number of substitutions over time t is the quantity

$$3\alpha t = -\frac{1}{4} \cdot \operatorname{trace}(Q) \cdot t = -\frac{1}{4} \cdot \log\det\bigl(P(t)\bigr). \tag{6.1}$$

This number is called the *branch length*. It can be computed from the substitution matrix $P(t)$ and is used to weight the edges in a phylogenetic X-tree.

One way to specify an *evolutionary model* is to give a phylogenetic X-tree T together with a rate matrix Q and an initial distribution for the root of T (which we here assume to be the stationary distribution on Ω). 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, and, according to the philosophy of algebraic statistics, we would like to regard it as an r-dimensional algebraic variety.

Such an algebraic representation does indeed exist. We shall explain it for the Jukes–Cantor DNA model on an X-tree T. Suppose that T has r edges and $|X| = n$ leaves. Let $P_i(t)$ denote the substitution matrix associated with the ith edge of the

tree. We write $3\alpha_i t_i = -\frac{1}{4}\log\det\bigl(P_i(t)\bigr)$ for the branch length of the ith edge, and we set $\pi_i = \frac{1}{4}(1 - e^{-4\alpha_i t_i})$ and $\theta_i = 1 - 3\pi_i$. Thus

$$P_i(t) = \begin{pmatrix} \theta_i & \pi_i & \pi_i & \pi_i \\ \pi_i & \theta_i & \pi_i & \pi_i \\ \pi_i & \pi_i & \theta_i & \pi_i \\ \pi_i & \pi_i & \pi_i & \theta_i \end{pmatrix}.$$

In algebraic geometry, we would regard θ_i and π_i as the homogeneous coordinates of a (complex) projective line $\mathbb{P}^1$, but in phylogenomics we limit our attention to the real segment specified by $\theta_i \geq 0$, $\pi_i \geq 0$, and $\theta_i + 3\pi_i = 1$.

Let Δ_{4^n-1} denote the set of all probability distributions on Ω^n. Since Ω^n has 4^n elements, namely, the DNA sequences of length n, the set Δ_{4^n-1} is a simplex of dimension $4^n - 1$. We identify the jth leaf of our tree T with the jth coordinate of a DNA sequence $(u_1, \ldots, u_n) \in \Omega^n$, and we introduce an unknown $p_{u_1 u_2 \cdots u_n}$ to represent the probability of observing the nucleotides $u_1, u_2, \ldots, u_n$ at the leaves $1, 2, \ldots, n$. The 4^n quantities $p_{u_1 u_2 \cdots u_n}$ are the coordinate functions on the simplex Δ_{4^n-1}, or, in the setting of algebraic geometry, on the projective space $\mathbb{P}^{4^n-1}$ obtained by complexifying Δ_{4^n-1}.

PROPOSITION 9. *In the Jukes–Cantor model on a tree T with r edges, the probability $p_{u_1 u_2 \cdots u_n}$ of making the observation $(u_1, u_2, \ldots, u_n) \in \Omega^n$ at the leaves is expressed as a polynomial which is multilinear of degree r in the model parameters (θ_1, π_1), $(\theta_2, \pi_2), \ldots, (\theta_n, \pi_n)$. Equivalently, in more geometric terms, the Jukes–Cantor model on T is the image of a multilinear map*

$$\phi \;:\; (\mathbb{P}^1)^r \;\longrightarrow\; \mathbb{P}^{4^n-1}. \tag{6.2}$$

The coordinates of the map ϕ are easily derived from the assumption that the substitution processes along different edges of T are independent. It turns out that the 4^n coordinates of ϕ are not all distinct. To see this, we work out the formulas explicitly for a very simple tree with three leaves.

EXAMPLE 10. *Let $n = r = 3$, and let T be the tree with three leaves, labeled by $X = \{1, 2, 3\}$, directly branching off the root of T. We consider the Jukes–Cantor DNA model with uniform root distribution on T. This model is a three-dimensional algebraic variety, given as the image of a trilinear map*

$$\phi \;:\; \mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \;\to\; \mathbb{P}^{63}.$$

The number of states in Ω^3 is $4^3 = 64$ but there are only five distinct polynomials occurring among the coordinates of the map ϕ. Let p_{123} be the probability of observing the same letter at all three leaves, p_{ij} the probability of observing the same letter at the leaves i, j and a different one at the third leaf, and p_{dis} the probability of seeing three distinct letters. Then

$$\begin{aligned}
p_{123} &= \theta_1\theta_2\theta_3 \;+\; 3\pi_1\pi_2\pi_3, \\
p_{dis} &= 6\theta_1\pi_2\pi_3 \;+\; 6\pi_1\theta_2\pi_3 \;+\; 6\pi_1\pi_2\theta_3 \;+\; 6\pi_1\pi_2\pi_3, \\
p_{12} &= 3\theta_1\theta_2\pi_3 \;+\; 3\pi_1\pi_2\theta_3 \;+\; 6\pi_1\pi_2\pi_3, \\
p_{13} &= 3\theta_1\pi_2\theta_3 \;+\; 3\pi_1\theta_2\pi_3 \;+\; 6\pi_1\pi_2\pi_3, \\
p_{23} &= 3\pi_1\theta_2\theta_3 \;+\; 3\theta_1\pi_2\pi_3 \;+\; 6\pi_1\pi_2\pi_3.
\end{aligned}$$

All 64 *coordinates of* ϕ *are given by these five trilinear polynomials, namely,*

$$\begin{aligned}
p_{AAA} = p_{CCC} = p_{GGG} = p_{TTT} &= \frac{1}{4} \cdot p_{123}, \\
p_{ACG} = p_{ACT} = \cdots = p_{GTC} &= \frac{1}{24} \cdot p_{dis}, \\
p_{AAC} = p_{AAT} = \cdots = p_{TTG} &= \frac{1}{12} \cdot p_{12}, \\
p_{ACA} = p_{ATA} = \cdots = p_{TGT} &= \frac{1}{12} \cdot p_{13}, \\
p_{CAA} = p_{TAA} = \cdots = p_{GTT} &= \frac{1}{12} \cdot p_{23}.
\end{aligned}$$

This means that our Jukes–Cantor model is the image of the simplified map

$$\phi' : \mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \to \mathbb{P}^4, \ \big((\theta_1,\pi_1),(\theta_2,\pi_2),(\theta_3,\pi_3)\big) \mapsto (p_{123}, p_{dis}, p_{12}, p_{13}, p_{23}).$$

In order to characterize the image of ϕ' *algebraically, we perform the following linear change of coordinates:*

$$\begin{aligned}
q_{111} &= p_{123} + \frac{1}{3}p_{dis} - \frac{1}{3}p_{12} - \frac{1}{3}p_{13} - \frac{1}{3}p_{23} = (\theta_1-\pi_1)(\theta_2-\pi_2)(\theta_3-\pi_3), \\
q_{110} &= p_{123} - \frac{1}{3}p_{dis} + p_{12} - \frac{1}{3}p_{13} - \frac{1}{3}p_{23} = (\theta_1-\pi_1)(\theta_2-\pi_2)(\theta_3+3\pi_3), \\
q_{101} &= p_{123} - \frac{1}{3}p_{dis} - \frac{1}{3}p_{12} + p_{13} - \frac{1}{3}p_{23} = (\theta_1-\pi_1)(\theta_2+3\pi_2)(\theta_3-\pi_3), \\
q_{011} &= p_{123} - \frac{1}{3}p_{dis} - \frac{1}{3}p_{12} - \frac{1}{3}p_{13} + p_{23} = (\theta_1+3\pi_1)(\theta_2-\pi_2)(\theta_3-\pi_3), \\
q_{000} &= p_{123} + p_{dis} + p_{12} + p_{13} + p_{23} = (\theta_1+3\pi_1)(\theta_2+3\pi_2)(\theta_3+3\pi_3).
\end{aligned}$$

This reveals that our model is the hypersurface in $\mathbb{P}^4$ *whose ideal equals*

$$I_T = \langle\, q_{000}q_{111}^2 - q_{011}q_{101}q_{110} \,\rangle.$$

If we set $\theta_i = 1 - 3\pi_i$, *then we get the additional constraint* $q_{000} = 1$.

The construction in this example generalizes to arbitrary trees T. There exists a change of coordinates, simultaneously on the *parameter space* $(\mathbb{P}^1)^r$ and on the *probability space* $\mathbb{P}^{4^n-1}$, such that the map ϕ in (6.2) becomes a monomial map in the new coordinates. This change of coordinates is known as the *Fourier transform* or as the *Hadamard conjugation* (see [25, 31, 57, 58]).

We regard the Jukes–Cantor DNA model on a tree T with n leaves and r edges as an algebraic variety of dimension r in $\mathbb{P}^{4^n-1}$, namely, it is the image of the map (6.2). Its homogeneous prime ideal I_T is generated by differences of monomials $q^a - q^b$ in the Fourier coordinates. In the phylogenetics literature (including the books [26, 50]), the polynomials in the ideal I_T are known as *phylogenetic invariants* of the model. The following result was shown in [57].

THEOREM 11. *The ideal* I_T *which defines the Jukes–Cantor model on a binary tree* T *is generated by monomial differences* $q^a - q^b$ *of degree at most three.*

It makes perfect sense to allow arbitrary distinct stochastic matrices $P(t)$ on the edges of the tree T. The resulting model is the *general Markov model* on the tree T. Allman and Rhodes [4, 5] determined the complete system of phylogenetic invariants for the general Markov model on a trivalent tree T.

An important problem in phylogenomics is that of identifying the maximum likelihood branch lengths, given a phylogenetic X-tree T, a rate matrix Q, and an alignment of sequences. For the Jukes–Cantor DNA model on three taxa, described in Example 10, the exact "analytic" solution of this optimization problem leads to an algebraic equation of degree 23. See [33, section 6] for details.

Let us instead consider the maximum likelihood estimation problem in the much simpler case of the Jukes–Cantor DNA model on two taxa. Here the tree T has only two leaves, labeled $X = \{1, 2\}$, directly branching off the root of T. The model is given by a surjective bilinear map

$$\phi \,:\, \mathbb{P}^1 \times \mathbb{P}^1 \,\to\, \mathbb{P}^1\,, \quad ((\theta_1, \pi_1), (\theta_2, \pi_2)) \,\mapsto\, (\, p_{12}, p_{dis}\,). \tag{6.3}$$

The coordinates of the map ϕ are

$$\begin{aligned} p_{12} &= \theta_1\theta_2 \,+\, 3\pi_1\pi_2, \\ p_{dis} &= 3\theta_1\pi_2 \,+\, 3\theta_2\pi_1 \,+\, 6\pi_1\pi_2. \end{aligned}$$

As before, we pass to affine coordinates by setting $\theta_i = 1 - 3\pi_i$ for $i = 1, 2$.

One crucial difference between the model (6.3) and Example 10 is that the parameters in (6.3) are *not identifiable*. Indeed, the inverse image of any point in $\mathbb{P}^1$ under the map ϕ is a curve in $\mathbb{P}^1 \times \mathbb{P}^1$. Suppose we are given data consisting of two aligned DNA sequences of length n where k of the bases are different. The corresponding point in $\mathbb{P}^1$ is $u = (n-k, k)$. The inverse image of u under the map ϕ is the curve in the affine plane with the equation

$$12n\pi_1\pi_2 \,-\, 3n\pi_1 \,-\, 3n\pi_2 \,+\, k = 0.$$

Every point (π_1, π_2) on this curve is an *exact fit* for the data $u = (n-k, k)$. Hence this curve equals the set of all maximum likelihood parameters for this model and the given data. We rewrite the equation of the curve as follows:

$$(1 - 4\pi_1)(1 - 4\pi_2) = 1 - \frac{4k}{3n}. \tag{6.4}$$

Recall from (6.1) that the branch length from the root to leaf i equals

$$3\alpha_i t_i = -\frac{1}{4} \cdot \log\det\bigl(P_i(t)\bigr) = -\frac{3}{4} \cdot \log(1 - 4\pi_i).$$

By taking logarithms on both sides of (6.4), we see that the curve of all maximum likelihood parameters becomes a line in the branch length coordinates:

$$3\alpha_1 t_1 \,+\, 3\alpha_2 t_2 = -\frac{3}{4} \cdot \log\left(1 - \frac{4k}{3n}\right). \tag{6.5}$$

The sum on the left-hand side is the distance from leaf 1 to leaf 2 in the tree T. Our discussion of the two-taxa model leads to the following formula which is known in evolutionary biology [26] under the name *Jukes–Cantor correction*.

PROPOSITION 12. *Given an alignment of two sequences of length n, with k differences between the bases, the ML estimate of the branch length equals*

$$\delta_{12} = -\frac{3}{4} \cdot \log\left(1 - \frac{4k}{3n}\right).$$

There has been recent progress on solving the likelihood equations exactly for small trees [15, 16, 33, 46]. We believe that these results will be useful in designing new algorithms for computing maximum likelihood branch lengths, and in better understanding the mathematical properties of existing methods (such as fastDNAml [40]) which are widely used by computational biologists.

It may also be the case that T is unknown, in which case the problem is not to select a point on a variety, but to select from (exponentially many) varieties. This problem is discussed in the next section.

The evolutionary models discussed above do not allow for insertion and deletion events. They also assume that sites evolve independently. Although many widely used models are based on these assumptions, biological reality calls for models that include insertion and deletion events [32], site interactions [52], and the flexibility to allow for genome dynamics such as rearrangements. Interested mathematicians will find a cornucopia of fascinating research problems arising from such more refined evolutionary models.

7. Phylogenetic Combinatorics. Fix a set X of n taxa. A *dissimilarity map* on X is a function $\delta : X \times X \to \mathbb{R}$ such that $\delta(x,x) = 0$ and $\delta(x,y) = \delta(y,x)$. The set of all dissimilarity maps on X is a real vector space of dimension $\binom{n}{2}$ which we identify with $\mathbb{R}^{\binom{n}{2}}$. A dissimilarity map δ is called a *metric on* X if the triangle inequality holds:

$$\delta(x,z) \;\leq\; \delta(x,y) + \delta(y,z) \qquad \text{for } x,y,z \in X.$$

The set of all metrics on X is a full-dimensional convex polyhedral cone in $\mathbb{R}^{\binom{n}{2}}$, called the *metric cone*. Phylogenetic combinatorics is concerned with the study of certain subsets of the metric cone which are relevant for biology. This field was pioneered in the 1980s by Andreas Dress and his collaborators; see Dress's 1998 ICM lecture [19] and the references given therein.

Let T be a phylogenetic X-tree whose edges have specified lengths. These lengths can be arbitrary nonnegative real numbers. The tree T defines a metric δ_T on X as follows: $\delta_T(x,y)$ equals the sum of the lengths of the edges on the unique path in T between the leaves labeled by x and y.

The *space of X-trees* is the following subset of the metric cone:

$$\mathcal{T}_X = \left\{ \delta_T \;:\; T \text{ is a phylogenetic } X\text{-tree} \right\} \subset \mathbb{R}^{\binom{n}{2}}. \tag{7.1}$$

Metric properties of the tree space $\mathcal{T}_X$ and its statistical and biological significance were studied by Billera, Holmes, and Vogtmann [9]. The following classical *four point condition* characterizes membership in the tree space.

THEOREM 13. *A metric δ on X lies in $\mathcal{T}_X$ if and only if, for any four taxa $u,v,x,y \in X$, $\delta(u,v) + \delta(x,y) \leq \max\{\delta(u,x) + \delta(v,y),\, \delta(u,y) + \delta(v,x)\}$.*

We refer to the book [50] for a proof of this theorem and several variants. To understand the structure of $\mathcal{T}_X$, let us fix the combinatorial type of a trivalent tree T. The number of choices of such trees is the *Schröder number*

$$(2n-5)!! = 1 \cdot 3 \cdot 5 \cdot \;\cdots\; \cdot (2n-7) \cdot (2n-5). \tag{7.2}$$

Since X has cardinality n, the tree T has $2n-3$ edges, and each of these edges corresponds to a *split* (A,B) of the set X into two nonempty disjoint subsets A and B. Let $\mathit{Splits}(T)$ denote the collection of all $2n-3$ splits (A,B) arising from T.

Each split (A, B) defines a *split metric* $\delta_{(A,B)}$ on X as follows:

$$\delta_{(A,B)}(x,y) = 0 \quad \text{if } (x \in A \text{ and } y \in A) \text{ or } (x \in B \text{ and } y \in B),$$
$$\delta_{(A,B)}(x,y) = 1 \quad \text{if } (x \in A \text{ and } y \in B) \text{ or } (y \in A \text{ and } x \in B).$$

The vectors $\{\delta_{(A,B)} : (A,B) \in \mathit{Splits}(T)\}$ are linearly independent in $\mathbb{R}^{\binom{n}{2}}$. Their nonnegative span is a cone $\mathcal{C}_T$ isomorphic to the orthant $\mathbb{R}^{2n-3}_{\geq 0}$.

PROPOSITION 14. *The space $\mathcal{T}_X$ of all X-trees is the union of the $(2n-5)!!$ orthants $\mathcal{C}_T$. It is hence a simplicial fan of pure dimension $2n-3$ in $\mathbb{R}^{\binom{n}{2}}$.*

The tree space $\mathcal{T}_X$ can be identified combinatorially with a simplicial complex of pure dimension $2n-4$, to be denoted $\widetilde{\mathcal{T}}_X$. The vertices of $\widetilde{\mathcal{T}}_X$ are the $2^{n-1}-1$ splits of the set X. We say that two splits (A, B) and (A', B') are *compatible* if at least one of the four sets $A \cap A'$, $A \cap B'$, $B \cap A'$, and $B \cap B'$ is the empty set. The following proposition is a combinatorial characterization of the tree space.

PROPOSITION 15. *A collection of splits of the set X forms a face in the simplicial complex $\widetilde{\mathcal{T}}_X$ if and only if that collection is pairwise compatible.*

The *phylogenetics problem* is to reconstruct a tree T from n aligned sequences. In principle, one can select from evolutionary models for all possible trees in order to find the maximum likelihood fit. Even if the maximum likelihood problem can be solved for each individual tree, this approach becomes infeasible in practice when n increases, because of the combinatorial explosion in the number (7.2) of trees. A number of alternative approaches have been suggested that attempt to find evolutionary models which fit *summaries* of the data. They build on the characterizations of trees given above.

Distance-based methods are based on the observation that trees can be encoded by metrics satisfying the four point condition (Theorem 13). Starting from a multiple sequence alignment, one can produce a dissimilarity map on the set X of taxa by computing the maximum likelihood distance between every pair of taxa, using Proposition 12. The resulting dissimilarity map δ is typically not a tree metric, i.e., it does not actually lie in the tree space $\mathcal{T}_X$. What needs to be done is to replace δ by a nearby tree metric $\delta_T \in \mathcal{T}_X$.

The method of choice for most biologists is the *neighbor-joining algorithm*, which provides an easy-to-compute map from the cone of all metrics onto $\mathcal{T}_X$. The algorithm is based on the following "cherry-picking theorem" [47, 56].

THEOREM 16. *Let δ be a tree metric on X. For every pair $i, j \in X$ set*

$$Q_\delta(i,j) = (n-2) \cdot \delta(i,j) - \sum_{k \neq i} \delta(i,k) - \sum_{k \neq j} \delta(j,k). \tag{7.3}$$

Then the pair $x, y \in X$ that minimizes $Q_\delta(x,y)$ is a cherry in the tree, i.e., x and y are separated by only one internal vertex z in the tree.

Neighbor-joining works as follows. Starting from an arbitrary metric δ on n taxa, one sets up the $n \times n$-matrix Q_δ whose (i,j)-entry is given by the formula (7.3), and one identifies the minimum off-diagonal entry $Q_\delta(x,y)$. If δ were a tree metric, then the internal vertex z which separates the leaves x and y would have the following distance from any other leaf k in the tree:

$$\delta(z,k) = \frac{1}{2}\big(\delta(x,k) + \delta(y,k) - \delta(x,y)\big). \tag{7.4}$$

One now removes the taxa x, y and replaces them by a new taxon z whose distance

to the remaining $n-2$ taxa is given by (7.4). This replaces the $n \times n$-matrix Q_δ by an $(n-1) \times (n-1)$ matrix, and one iterates the process.

This neighbor-joining algorithm recursively constructs a tree T whose metric δ_T is reasonably close to the given metric δ. If δ is a tree metric, then the method is guaranteed to reconstruct the correct tree. More generally, instead of estimating pairwise distances, one can attempt to (more accurately) estimate the sum of the branch lengths of subtrees of size $m \geq 3$.

We define an *m-dissimilarity map* on X to be a function $\delta : X^m \to \mathbb{R}$ such that $\delta(i_1, i_2, \ldots, i_m) = \delta(i_{\pi(1)}, i_{\pi(2)}, \ldots, i_{\pi(m)})$ for all permutations π on $\{1, \ldots, m\}$ and $\delta(i_1, i_2, \ldots, i_m) = 0$ if the taxa $i_1, i_2, \ldots, i_m$ are not distinct. The set of all m-dissimilarity maps on X is a real vector space of dimension $\binom{n}{m}$ which we identify with $\mathbb{R}^{\binom{n}{m}}$. Every X-tree T gives rise to an m-dissimilarity map δ_T as follows. We define $\delta_T(i_1, \ldots, i_m)$ to be the sum of all branch lengths in the subtree of T spanned by $i_1, \ldots, i_m \in X$.

The following theorem [17, 41] is a generalization of Theorem 16. It leads to a generalized neighbor-joining algorithm which provides a better approximation of the maximum likelihood tree and parameters.

THEOREM 17. *Let T be an X-tree and $m < n = |X|$. For any $i, j \in X$ set*

$$Q_T(i,j) \;=\; \binom{n-2}{m-1} \sum_{Y \in \binom{X \setminus \{i,j\}}{m-2}} \delta_T(i,j,Y) \;-\; \sum_{Y \in \binom{X \setminus \{i\}}{m-1}} \delta_T(i,Y) \;-\; \sum_{Y \in \binom{X \setminus \{j\}}{m-1}} \delta_T(j,Y).$$

Then the pair $x, y \in X$ that minimizes $Q_T(x,y)$ is a cherry in the tree T.

The subset of $\mathbb{R}^{\binom{n}{m}}$ consisting of all m-dissimilarity maps δ_T arising from trees T is a polyhedral space which is the image of the tree space $\mathcal{T}_X$ under a piecewise-linear map $\mathbb{R}^{\binom{n}{2}} \to \mathbb{R}^{\binom{n}{m}}$. We do not know a simple characterization of this m-version of tree space which extends the four point condition.

Here is another natural generalization of the space of trees. Fix an m-dissimilarity map $\delta : X^m \to \mathbb{R}$ and consider any $(m-2)$-element subset $Y \in \binom{X}{m-2}$. We get an induced dissimilarity map $\delta\!/_Y$ on $X \backslash Y$ by setting

$$\delta\!/_Y(i,j) = \delta(i,j,Y) \qquad \text{for all } i, j \in X \backslash Y.$$

We say that δ is an *m-tree* if $\delta\!/_Y$ is a tree metric for all $Y \in \binom{X}{m-2}$. Thus, by Theorem 13, an m-dissimilarity map δ on X is an m-tree if

$$\delta(i,j,Y) + \delta(k,l,Y) \;\leq\; \max\{\delta(i,k,Y) + \delta(j,l,Y),\, \delta(i,l,Y) + \delta(k,j,Y)\}$$

for all $Y \in \binom{X}{m-2}$ and all $i, j, k, l \in X \backslash Y$.

Let $Gr_{m,n}$ denote the subset of $\mathbb{R}^{\binom{n}{m}}$ consisting of all m-trees. The space $Gr_{m,n}$ is a polyhedral fan which is slightly larger than the *tropical Grassmannian* studied in [54]. For every m-tree $\delta \in Gr_{m,n}$ there is an $(m-1)$-dimensional tree-like space whose "leaves" are the taxa in X. This is the *tropical linear space* defined in [53]. This construction, which is described in [54, section 6] and [44, section 3.5], specializes to the construction of an X-tree T from its metric δ_T when $m = 2$. The study of m-trees and the tropical Grassmannian was anticipated in [19, 20]. The Dress–Wenzel theory of *matroids with coefficients* [20] contains our m-trees as a special case. The space $Gr_{m,n}$ of all m-trees is discussed in the context of buildings in [19]. Note that the tree space $\mathcal{T}_X$ in (7.1) is precisely the tropical Grassmannian $Gr_{2,n}$.

It is an open problem to find a natural and easy-to-compute projection from $\mathbb{R}^{\binom{n}{m}}$ onto $Gr_{m,n}$ which generalizes the neighbor-joining method. Such a variant of neighbor-joining would be likely to have applications for more intricate biological data that are not easily explained by a tree model. We close this section by discussing an example.

EXAMPLE 18. *Fix a set of six taxa, $X = \{1,2,3,4,5,6\}$, and let $m = 3$. The space of 3-dissimilarity maps on X is identified with $\mathbb{R}^{20}$. An element $\delta \in \mathbb{R}^{20}$ is a 3-tree if $\delta/_i$ is a tree metric on $X\backslash\{i\}$ for all i. Equivalently,*

$$\delta(i,j,k) + \delta(i,l,m) \;\leq\; \max\bigl\{\delta(i,j,l) + \delta(i,k,m),\, \delta(i,j,m) + \delta(i,k,l)\bigr\}$$

for all $i,j,k,l,m \in X$. The set $Gr_{3,6}$ of all 3-trees is a 10-dimensional polyhedral fan. Each cone in this fan contains the 6-dimensional linear space L consisting of all 3-dissimilarity maps of the particular form

$$\delta(i,j,k) \;=\; \omega_i + \omega_j + \omega_k \qquad \text{for some } \omega \in \mathbb{R}^6.$$

The quotient $Gr_{3,6}/L$ is a 4-dimensional fan in the 14-dimensional real vector space $\mathbb{R}^{20}/L$. Let $\tilde{G}r_{3,6}$ denote the 3-dimensional polyhedral complex obtained by intersecting $Gr_{3,6}/L$ with a sphere around the origin in $\mathbb{R}^{20}/L$.

It was shown in [54, section 5] *that $\tilde{G}r_{3,6}$ is a 3-dimensional simplicial complex consisting of 65 vertices, 550 edges, 1,395 triangles, and 1,035 tetrahedra. Each of the 1,035 tetrahedra parameterizes 6-tuples of tree metrics*

$$\bigl(\,\delta/_1,\,\delta/_2,\,\delta/_3,\,\delta/_4,\,\delta/_5,\,\delta/_6\,\bigr),$$

where the tree topologies on five taxa are fixed. The homology of the tropical Grassmannian $\tilde{G}r_{3,6}$ is concentrated in the top dimension and is free abelian:

$$H_3\bigl(\tilde{G}r_{3,6},\mathbb{Z}\bigr) = \mathbb{Z}^{126}.$$

If T is an X-tree and δ_T the corresponding 3-dissimilarity map (as in Theorem 17), then it is easy to check that δ_T lies in $Gr_{3,6}$. The set of all 3-trees of the special form $\delta = \delta_T$ has codimension 1 in $Gr_{3,6}$. It is the intersection of $Gr_{3,6}$ with the 15-dimensional linear subspace of $\mathbb{R}^{20}$ defined by the equations

$$\begin{aligned}
&\delta(123)+\delta(145)+\delta(246)+\delta(356) = \delta(124)+\delta(135)+\delta(236)+\delta(456),\\
&\delta(123)+\delta(145)+\delta(346)+\delta(256) = \delta(134)+\delta(125)+\delta(236)+\delta(456),\\
&\delta(123)+\delta(245)+\delta(146)+\delta(356) = \delta(124)+\delta(235)+\delta(136)+\delta(456),\\
&\delta(123)+\delta(345)+\delta(246)+\delta(156) = \delta(234)+\delta(135)+\delta(126)+\delta(456),\\
&\delta(123)+\delta(345)+\delta(146)+\delta(256) = \delta(134)+\delta(235)+\delta(126)+\delta(456).
\end{aligned}$$

Working modulo L and intersecting with a suitable sphere, the tree space $\widetilde{\mathcal{T}}_X$ is a 2-dimensional simplicial complex, consisting of $105 = 5!!$ triangles. To be precise, the simplicial complex in Proposition 15 is the join of this triangulated surface with the 5-simplex on X. Theorem 17 relates to the following geometric picture: the triangulated surface $\widetilde{\mathcal{T}}_X$ sits inside the triangulated threefold $\tilde{G}r_{3,6}$, namely, as the solution set of the five equations.

8. Back to the Data. In section 2, a conjecture was proposed based on our finding that the "Meaning of Life" sequence (2.1) is present (without mutations, insertions, or deletions) in orthologous regions in ten vertebrate genomes. In this section we explain how the various ideas outlined throughout this paper can be used to estimate the probability that such an extraordinary degree of conservation would occur by chance. The mechanics of the calculation also provide a glimpse into the types of processing and analyses that are performed in computational biology. Two research papers dealing with this subject matter are [7, 21].

What we shall compute in this section is the probability under the Jukes–Cantor model that a single ancestral base that is not under selection (and is therefore free to mutate) is identical in the ten present day vertebrates.

Step 1 (genomes). The National Center for Biotechnology Information (NCBI—http://www.ncbi.nlm.nih.gov/) maintains a public database called GENBANK which contains all publicly available genome sequences from around the world. Large sequencing centers that receive public funding are generally required to deposit raw sequences into this database within 24 hours of processing by sequencing machines, and thus many automatic pipelines have been set up for generating and depositing sequences. The growth in GENBANK has been spectacular. The database contained only 680,000 base pairs when it was started in 1982, and this number went up to 49 million by 1990. There are currently 44 billion base pairs of DNA in GENBANK.

The ten genomes of interest are not all complete, but are all downloadable from GENBANK, either in pieces mapped to chromosomes (e.g., for human) or as collections of subsequences called *contigs* (for less complete genomes).

Step 2 (annotation). In order to answer our question we need to know where genes are in the genomes. Some genomes have annotations that were derived experimentally, but *all* the genomes are annotated using HMMs (section 4) shortly after the release of the sequence. These annotations are performed by centers such as at UC Santa Cruz (http://genome.ucsc.edu/) as well as by individual authors of programs. It remains an open problem to accurately annotate genomes. But HMM programs are quite good on average. For example, typically 98% of coding bases are predicted correctly to be in genes. On the other hand, boundaries of exons are often misannotated: current state of the art methods only achieve accuracies of about 80% [6].

Step 3 (alignment). We start out by performing a genome alignment. Current methods for aligning whole genomes are all based, to varying degrees, on the pair HMM ideas of section 5. Although in practice it is not possible to align sequences containing billions or even millions of base pairs with HMMs, pair HMMs are subroutines of more complex alignment strategies where smaller regions for alignment are initially identified from the entire genomes by fast string matching algorithms [10]. The ten vertebrate whole genome alignments which gave rise to Conjecture 1 are accessible at http://bio.math.berkeley.edu/genomes/.

Step 4 (finding neutral DNA). In order to compute the probability that a certain subsequence is conserved between genomes, it is necessary to estimate the *neutral rate of evolution*. This is done by estimating parameters for an evolutionary model of base pairs in the genome that are not under selection, and are therefore free to mutate. Since neutral regions are difficult to identify a priori, commonly used surrogates are synonymous substitutions in codons (section 3). Because synonymous substitutions do not change the amino acids, it is unlikely that they are selected for or against, and various studies have shown that such data provide good estimates for neutral

Table 8.1 *Jukes–Cantor pairwise distance estimates.*

	gg	hs	mm	pt	rn	cf	dr	tn	tr	xt
gg	–	0.831	0.928	0.831	0.925	0.847	1.321	1.326	1.314	1.121
hs	–	–	0.414	0.013	0.411	0.275	1.296	1.274	1.290	1.166
mm	–	–	–	0.413	0.176	0.441	1.256	1.233	1.264	1.218
pt	–	–	–	–	0.411	0.275	1.291	1.267	1.288	1.160
rn	–	–	–	–	–	0.443	1.255	1.233	1.258	1.212
cf	–	–	–	–	–	–	1.300	1.251	1.269	1.154
dr	–	–	–	–	–	–	–	1.056	1.067	1.348
tn	–	–	–	–	–	–	–	–	0.315	1.456
tr	–	–	–	–	–	–	–	–	–	1.437

mutation rates. By searching through the annotations and alignments, we identified $n = 14{,}202$ fourfold degenerate sites. These can be used for analyzing probabilities of neutral mutations.

Step 5 (deriving a metric). We would ideally like to use maximum likelihood techniques to reconstruct a tree T with branch lengths from the alignments of the four-fold degenerate sites. One approach is to try to use a maximum likelihood approach, but this is difficult to do reliably because of the complexity of the likelihood equations, even for the Jukes–Cantor models with $|X| = 10$. An alternative approach is to estimate pairwise distances between species i, j using the formula in Proposition 12. The resulting metric on the set $X = \{\mathrm{gg}, \mathrm{hs}, \mathrm{mm}, \mathrm{pt}, \mathrm{rn}, \mathrm{cf}, \mathrm{dr}, \mathrm{tn}, \mathrm{tr}, \mathrm{xt}\}$ is given in Table 8.1. For example, the pairwise alignment between human and chicken (extracted from the multiple alignment) has $n = 14{,}202$ positions, of which $k = 7{,}132$ are different. Thus, the Jukes–Cantor distance between the genomes of human and chicken equals

$$-\frac{3}{4} \cdot \log\left(1 - \frac{4k}{3n}\right) = -\frac{3}{4} \cdot \log\left(\frac{14078}{42606}\right) = 0.830536\ldots.$$

Step 6 (building a tree). From the pairwise distances in Table 8.1 we construct a phylogenetic X-tree using the neighbor-joining algorithm (section 7). The tree with the inferred branch lengths is shown in Figure 8.1. The tree is drawn such that the branch lengths are consistent with the horizontal distances in the diagram. The root of the tree was added manually in order to properly indicate the ancestral relationships between the species.

At this point we wish to add a philosophical remark: *The tree in Figure* 8.1 *is a point on an algebraic variety!* Indeed, that variety is the Jukes–Cantor model (Proposition 9), and the preimage coordinates (θ_i, π_i) of that point are obtained by exponentiating the branch lengths as described in section 6.

Step 7 (calculating the probability). We are now given a specific point on the variety representing the Jukes–Cantor model on the tree depicted in Figure 8.1. Recall from Proposition 9 that this variety, and hence our point, lives in a projective space of dimension $4^{10} - 1 = 1{,}048{,}575$. What we are interested in are four specific coordinates of that point, namely, the probabilities that the same nucleotide occurs in every species:

$$p_{AAAAAAAAAA} = p_{CCCCCCCCCC} = p_{GGGGGGGGGG} = p_{TTTTTTTTTT}. \tag{8.1}$$

As discussed in section 6, this expression is a multilinear polynomial in the edge parameters (θ_i, ϕ_i). When we evaluate it at the parameters derived from the branch

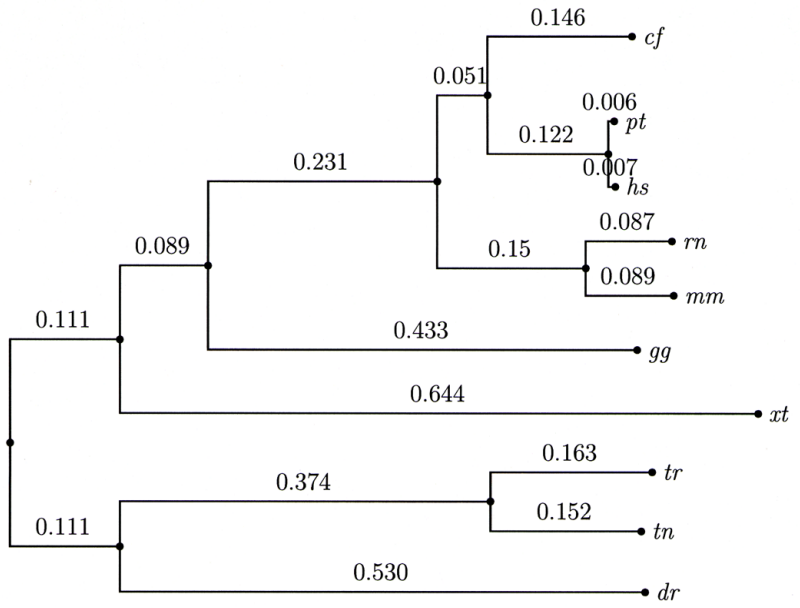

Fig. 8.1 *Neighbor-joining tree from alignment of codons in ten vertebrates.*

lengths in Figure 8.1, we find that

$$p_{AAAAAAAAAA} = 0.009651\ldots.$$

Returning to the "Meaning of Life" sequence (2.1), this implies the following.

PROPOSITION 19. *Assuming the probability distribution on* Ω^{10} *given by the Jukes–Cantor model on the tree in Figure* 8.1, *the probability of observing a sequence of length* 42 *unchanged at a given location in the ten vertebrate genomes within a neutrally evolving region equals* $(0.038604)^{42} = 4.3 \cdot 10^{-60}$.

This calculation did not take into account the fact that the "Meaning of Life" sequence may occur in an arbitrary location of the genome in question. In order to adjust for this, we can multiply the number in Proposition 19 by the length of the genomes. The human genome contains approximately 2.8 billion nucleotides, so it is reasonable to conclude that the probability of observing a sequence of length 42 unchanged *somewhere* in the ten vertebrate genomes is approximately

$$2.8 \cdot 10^9 \ \times \ 4.3 \cdot 10^{-60} \simeq 10^{-50}.$$

This probability is a very small number, i.e., it is unlikely that the remarkable properties of the sequence (2.1) occurred by "chance." Despite the shortcomings of the Jukes–Cantor model discussed at the end of section 6, we believe that Proposition 19 constitutes a sound argument in support of Conjecture 1.

Acknowledgments. The vertebrate whole genome alignments we have analyzed were assembled by Nicolas Bray and Colin Dewey. We also thank Sourav Chatterji and Von Bing Yap for their help in searching through the alignments.

REFERENCES

[1] D. N. Adams, *The Hitchhiker's Guide to the Galaxy*, Pan Books, London, 1979.

[2] M. Alexandersson, N. Bray, and L. Pachter, *Pair hidden Markov models*, in Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics, L. B. Jorde, P. Little, M. Dunn, and S. Subramaniam, eds., Wiley-Interscience, New York, 2005; available online through Wiley Interscience.

[3] M. Alexandersson, S. Cawley, and L. Pachter, *SLAM—Cross-species gene finding and alignment with a generalized pair hidden Markov model*, Genome Res., 13 (2003), pp. 496–502.

[4] E. Allman and J. Rhodes, *Phylogenetic invariants for the general Markov model of sequence mutation*, Math. Biosci., 186 (2003), pp. 133–144.

[5] E. Allman and J. Rhodes, *Phylogenetic Ideals and Varieties for the General Markov Model*, preprint, http://www.citebase.org/abstract?id=oai:arXiv.org:math/0410604, 2004.

[6] J. Ashurst and J. E. Collins, *Gene annotation: Prediction and testing*, Ann. Rev. Genomics Human Genetics, 4 (2003), pp. 69–88.

[7] G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler, *Ultraconserved elements in the human genome*, Science, 304 (2004), pp. 1321–1325.

[8] P. Bickel and K. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, San Francisco, CA, 1976.

[9] L. Billera, S. Holmes, and K. Vogtmann, *Geometry of the space of phylogenetic trees*, Adv. Appl. Math., 27 (2001), pp. 733–767.

[10] N. Bray and L. Pachter, *MAVID: Constrained ancestral alignment of multiple sequences*, Genome Res., 14 (2004), pp. 693–699.

[11] P. Bucher and K. Hofmann, *A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system*, in Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology (ISMB '96), AAAI Press, Menlo Park, CA, 1996, pp. 44–51.

[12] C. Burge and S. Karlin, *Prediction of complete gene structures in human genomic DNA*, J. Mol. Biol., 268 (1997), pp. 78–94.

[13] A. Campbell, J. Mrazek, and S. Karlin, *Genome signature comparisons among prokaryote, plasmid and mitochondrial DNA*, Proc. Natl. Acad. Sci. USA, 96 (1999), pp. 9184–9189.

[14] S. Chatterji and L. Pachter, *Multiple organism gene finding by collapsed Gibbs sampling*, in Proceedings of the Eighth Annual International Conference on Computational Molecular Biology—RECOMB 2004, San Diego, CA, ACM, 2004, pp. 187–193.

[15] B. Chor, M. Hendy, and S. Snir, *Maximum likelihood Jukes-Cantor triplets: Analytic solutions*, Mol. Biol. Evol., 23 (2006), pp. 626–632.

[16] B. Chor, A. Khetan, and S. Snir, *Maximum likelihood on four taxa phylogenetic trees: Analytic solutions*, in Proceedings of the Seventh Annual Conference on Research in Computational Molecular Biology—RECOMB 2003, Berlin, ACM, 2003, pp. 76–83.

[17] M. Contois and D. Levy, *Small trees and generalized neighbor-joining*, in Algebraic Statistics for Computational Biology, L. Pachter and B. Sturmfels, eds., Cambridge University Press, Cambridge, UK, 2005, pp. 335–346.

[18] R. Durbin, S. R. Eddy, A. Korgh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK, 1999.

[19] A. Dress and W. Terhalle, *The tree of life and other affine buildings*, in Proceedings of the International Congress of Mathematicians, Vol. III (Berlin, 1998), Doc. Math., 1998, pp. 565–574.

[20] A. Dress and W. Wenzel, *Grassmann-Plücker relations and matroids with coefficients*, Adv. Math., 86 (1991), pp. 68–110.

[21] M. Drton, N. Eriksson, and G. Leung, *Ultra-conserved elements in vertebrate and fly genomes*, in Algebraic Statistics for Computational Biology, L. Pachter and B. Sturmfels, eds., Cambridge University Press, Cambridge, UK, 2005, pp. 387–402.

[22] J. Eisen, *Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis*, Genome Res., 8 (1998), pp. 163–167.

[23] E. E. Eichler and D. Sankoff, *Structural dynamics of eukaryotic chromosome evolution*, Science, 301 (2003), pp. 793–797.

[24] S. Elizalde, *Inference functions*, in Algebraic Statistics for Computational Biology, L. Pachter and B. Sturmfels, eds., Cambridge University Press, Cambridge, UK, 2005, pp. 215–225.

[25] S. Evans and T. Speed, *Invariants of some probability models used in phylogenetic inference*, Ann. Statist., 21 (1993), pp. 355–377.

[26] J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, Sunderland, MA, 2003.
[27] D. Fernández-Baca and B. Venkatachalam, *Parametric Sequence Alignment*, Handbook on Computational Molecular Biology, S. Aluru, ed., Chapman and Hall/CRC Press, Boca Raton, FL, 2005.
[28] S. E. Fienberg, *The Analysis of Cross-Classified Categorical Data*, 2nd ed., MIT Press, Cambridge, MA, 1980.
[29] D. Gusfield, *Algorithms on Strings, Trees and Sequences*, Cambridge University Press, Cambridge, UK, 1997.
[30] I. Hallgrímsdóttir, R. A. Milowski, and J. Yu, *The EM algorithm for hidden Markov models*, in Algebraic Statistics for Computational Biology, L. Pachter and B. Sturmfels, eds., Cambridge University Press, Cambridge, UK, 2005, pp. 250–263.
[31] M. D. Hendy and D. Penny, *Spectral analysis of phylogenetic data*, J. Classification, 10 (1993), pp. 5–24.
[32] I. Holmes and W. J. Bruno, *Evolutionary HMMs: A Bayesian approach to multiple alignment*, Bioinform., 17 (2001), pp. 803–820.
[33] S. Hoşten, A. Khetan, and B. Sturmfels, *Solving the likelihood equations*, Found. Comput. Math., 5 (2005), pp. 389–407.
[34] M. Kac, G.-C. Rota, and J. T. Schwartz, *Discrete Thoughts*, Birkhäuser, Boston, 1986.
[35] R. M. Karp, *Mathematical challenges from genomics and molecular biology*, Notices Amer. Math. Soc., 49 (2002), pp. 544–553.
[36] M. Kellis, B. Birren, and E. Lander, *Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae*, Nature, 8 (2004), pp. 617–624.
[37] D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman, *A generalized hidden Markov model for the recognition of human genes in DNA*, in Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology (ISMB '96), AAAI Press, Menlo Park, CA, 1996, pp. 134–142.
[38] E. S. Lander et al., *Initial sequencing and analysis of the human genome*, Nature, 409 (2001), pp. 860–921.
[39] E. M. Myers et al., *A whole-genome assembly of Drosophila*, Science, 287 (2000), pp. 2196–2204.
[40] G. J. Olsen, H. Matsuda, R. Hagstrom, and R. Overbeek, *fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood*, Comput. Appl. Biosci., 10 (1994), pp. 41–48
[41] L. Pachter and D. Speyer, *Reconstructing trees from subtree weights*, Appl. Math. Lett., 17 (2004), pp. 615–621.
[42] L. Pachter and B. Sturmfels, *Tropical geometry of statistical models*, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 16132–16137
[43] L. Pachter and B. Sturmfels, *Parametric inference for biological sequence analysis*, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 16138–16143.
[44] L. Pachter and B. Sturmfels, eds., *Algebraic Statistics for Computational Biology*, Cambridge University Press, Cambridge, UK, 2005.
[45] P. Pevzner and G. Tesler, *Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 7672–7677.
[46] R. Sainudiin and R. Yoshida, *Applications of interval methods to phylogenetics*, in Algebraic Statistics for Computational Biology, L. Pachter and B. Sturmfels, eds., Cambridge University Press, Cambridge, UK, 2005, pp. 359–374.
[47] N. Saitou and M. Nei, *The neighbor joining method: A new method for reconstructing phylogenetic trees*, Mol. Bio. Evol., 4 (1987), pp. 406–425.
[48] A. Sandelin, P. Bailey, S. Bruce, P. Engström, J. M. Klos, W. W. Wasserman, J. Ericson, and B. Lenhard, *Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes*, BMC Genomics, 5 (2004), p. 99.
[49] D. Sankoff and J. H. Nadeau, *Chromosome rearrangements in evolution: From gene order to genome sequence and back*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 11188–11189.
[50] C. Semple and M. Steel, *Phylogenetics*. Oxford University Press, Oxford, UK, 2003.
[51] A. Siepel et al., *Evolutionarily conserved elements in vertebrate, insect, worm and yeast genomes*, Genome Res., 15 (2005), pp. 1034–1050.
[52] A. Siepel and D. Haussler, *Phylogenetic estimation of context-dependent substitution rates by maximum likelihood*, Mol. Bio. Evol., 21 (2004), pp. 468–488.
[53] D. Speyer, *Tropical Linear Spaces*, preprint, http://arxiv.org/abs/math/0410455.
[54] D. Speyer and B. Sturmfels, *The tropical Grassmannian*, Adv. Geom., 4 (2004), pp. 389–411.
[55] R. P. Stanley, *Enumerative Combinatorics*, Vol. 1, Cambridge Stud. Adv. Math. 49, Cambridge University Press, Cambridge, UK, 1997.

[56] J. A. Studier and K. J. Keppler, *A note on the neighbor-joining method of Saitou and Nei*, Mol. Bio. Evol., 5(1988), pp. 729–731.
[57] B. Sturmfels and S. Sullivant, *Toric ideals of phylogenetic invariants*, J. Comput. Biol., 12 (2005), pp. 204–228.
[58] L. Székely, M. Steel, and P. Erdös, *Fourier calculus on evolutionary trees*, Adv. Appl. Math., 14 (1993), pp. 200–210.
[59] J. C. Venter et al., *The sequence of the human genome*, Science, 291 (2001), pp. 1304–1351.
[60] M. Waterman, M. Eggert, and E. Lander, *Parametric sequence comparisons*, Proc. Natl. Acad. Sci. USA, 89 (1992), pp. 6090–6093.
[61] J. Watson and F. Crick, *A structure for deoxyribose nucleic acid*, Nature, 171 (1953), pp. 964–967.
[62] A. Woolfe et al., *Highly conserved non-coding sequences are associated with vertebrate development*, PLoS Biology, 3 (2005), pp. 116–130.
[63] V. B. Yap and L. Pachter, *Identification of evolutionary hotspots in the rodent genomes*, Genome Res., 14 (2004), pp. 574–579.