Tropical geometry of statistical models

Lior Pachter and Bernd Sturmfels[†]

Department of Mathematics, University of California, Berkeley, CA 94720

Communicated by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, September 10, 2004 (received for review January 25, 2004)

This article presents a unified mathematical framework for inference in graphical models, building on the observation that graphical models are algebraic varieties. From this geometric viewpoint, observations generated from a model are coordinates of a point in the variety, and the sum–product algorithm is an efficient tool for evaluating specific coordinates. Here, we address the question of how the solutions to various inference problems depend on the model parameters. The proposed answer is expressed in terms of tropical algebraic geometry. The Newton polytope of a statistical model plays a key role. Our results are applied to the hidden Markov model and the general Markov model on a binary tree.

This article presents a unified mathematical framework for probabilistic inference with statistical models, such as graphical models. Our approach is summarized by the following theses:

(*i*) Statistical models are algebraic varieties.
(*ii*) Every algebraic variety can be tropicalized.
(*iii*) Tropicalized statistical models are fundamental for parametric inference.

1. Algebraic Statistics, Tropical Geometry, and Inference

By a *statistical model*, we mean a family of joint probability distributions for a collection of discrete random variables $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$. Thesis *i* states that many families of interest can be characterized by polynomials in the joint probabilities $p_{\sigma_1 \cdots \sigma_n} = \mathrm{Prob}(Y_1 = \sigma_1, \ldots, Y_n = \sigma_n)$. Although the variety defined by these polynomials contains points that are not in the model (for example, points with negative coordinates), the emerging field of algebraic statistics (1, 2) offers practical and useful algorithms for studying statistical models.

Tropicalization means replacing the arithmetic operations $(+, \times)$ by the operations $(\min, +)$. This process captures the essence of what happens when the joint probabilities $p_{\sigma_1 \cdots \sigma_n}$ are replaced by their logarithms. The tropicalization of an algebraic variety is a piecewise-linear set that has many features familiar from algebraic geometry (3, 4). In particular, the tropicalization of a statistical model is a piecewise-linear set in the space with logarithmic coordinates $-\log(p_{\sigma_1 \cdots \sigma_n})$.

Thesis *iii* states that tropical algebraic geometry of statistical models is of fundamental interest in analyzing the behavior of inference algorithms under the variation of model parameters. By *inference*, we mean the evaluation of one or more coordinates of a single point on the algebraic variety, in either $(+, \times)$ or $(\min, +)$ arithmetic. This evaluation corresponds to a form of inference that is used for graphical models in statistical learning theory (5), but it differs from other (more classical) notions of inference in mathematical statistics. By *parametric inference*, we mean the analysis of the dependence of inference on parameters.

To give a more concrete discussion of parametric inference, it is useful to focus on directed graphical models. A *directed graphical model* (or *Bayesian network*) is a finite directed acyclic graph G with two kinds of vertices (*observed variables* $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ and *hidden variables* $\mathbf{X} = \{X_1, \ldots, X_m\}$), where each edge is labeled by a transition matrix whose entries are linear forms in some parameters. The rules of discrete probability express the observed probabilities $p_{\sigma_1 \cdots \sigma_n}$ as polynomials of a degree $\leq E$ in the parameters, where E is the number of edges of G. The polynomials parameterize the graphical model as an algebraic variety.

The following are two types of inference questions from statistical learning theory for graphical models.

1. The calculation of *marginal probabilities*:

$$p_{\sigma_1 \cdots \sigma_n} = \sum_{h_1, \ldots, h_m} \mathrm{Prob}(X_1 = h_1, \ldots, X_m = h_m, \\ Y_1 = \sigma_1, \ldots, Y_n = \sigma_n), \quad \text{and}$$

2. The calculation of *maximum a posteriori* (*MAP*) log probabilities:

$$\delta_{\sigma_1 \cdots \sigma_n} = \min_{h_1, \ldots, h_m} -\log(\mathrm{Prob}(X_1 = h_1, \ldots, X_m = h_m, \\ Y_1 = \sigma_1, \ldots, Y_n = \sigma_n)),$$

where the h_i range over all of the possible assignments for the hidden random variables X_i. Together, these two primitives can be used effectively to solve a range of other statistical learning inference problems, including the calculation of conditional probabilities and other quantities of interest. The key to these statistical learning inference questions for graphical models is the *sum–product algorithm* (6), which is also known as the *generalized distributive law* (7). This polynomial-time algorithm (in the case that the graph has constant clique size) is used, both in ordinary arithmetic $(+, \times)$ and in tropical arithmetic $(\min, +)$, to solve problems 1 and 2 *efficiently*. For more background on the sum–product algorithm, and for connections to message passing and the junction tree algorithm, see ref. 5.

Although the sum–product algorithm provides efficient solutions to the basic inference problems 1 and 2, it only applies to one coordinate $p_{\sigma_1 \cdots \sigma_n}$ of one distribution at a time. We are interested in the *parametric* versions of the inference problems. They can be phrased as follows:

3. Find all parameter values for a model that result in the same values for all $p_{\sigma_1 \cdots \sigma_n}$.
4. Given observations $\mathbf{Y} = \sigma$ and hidden data $\mathbf{X} = \mathbf{h}$, identify all parameter values such that $\mathbf{h}$ is the most likely explanation for the observations σ.

As we will show, the following *modeling* problems are related fundamentally to problems 3 and 4:

5. Which (parameter-independent) relations on the probabilities $p_{\sigma_1 \cdots \sigma_n}$ does the model imply?
6. Describe the tropicalization of the variety that corresponds to a graphical model.

Problem 5 asks for the ideal of *polynomial invariants* of a statistical model (1). Invariants have been investigated in phylogenetics (8, 9), where they can help to identify good trees for aligned DNA sequences.

The primary goal of our study is to give a practical answer to problem 4 for graphical models. Our main algorithmic result is an efficient procedure for parametric inference that can be viewed as a polytopal analog of the sum–product algorithm.

Abbreviations: MAP, maximum *a posteriori*; HMM, hidden Markov model.

[†] To whom correspondence should be addressed. E-mail: bernd@math.berkeley.edu.

16132–16137 | PNAS | November 16, 2004 | vol. 101 | no. 46

www.pnas.org/cgi/doi/10.1073/pnas.0406010101

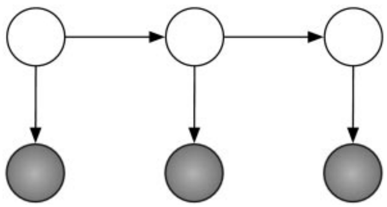

Fig. 1. The HMM of length 3.

The efficiency is based on the complexity estimates for Newton polytopes that we derive in section 4. The resulting *polytope propagation algorithm* is applied to problems in biological sequence analysis in ref. 10.

The mathematics developed in sections 3 and 4 is of independent interest. It also furnishes tools for parametric inference (problems 3 and 4) and parametric modeling (problems 5 and 6), which are applicable to a wide range of statistical problems. We demonstrate this point of view by analyzing the *hidden Markov model* (HMM) and the general Markov model on a binary tree in sections 2 and 5, respectively.

2. Algebraic Representation of HMMs

A graphical model is an algebraic variety that is presented as the image of a highly structured polynomial map f: $\mathbf{R}^d \to \mathbf{R}^m$. Here, $\mathbf{R}^d$ is the space in which coordinates are the model parameters $s_1, \ldots, s_d$, and $\mathbf{R}^m$ is the space in which coordinates $p_\sigma = p_{\sigma_1 \cdots \sigma_n}$ are the joint probabilities for the observed random variables. In applications, the integer m is much larger than the integer d; in fact, it is so large that one can only look at one coordinate p_σ at a time. Each coordinate $f_\sigma = f_\sigma(s_1, \ldots, s_d)$ of the map f is a polynomial function in $s_1, \ldots, s_d$. The efficient evaluation of these functions relies on the sum–product algorithm. Here, we study the (parametric) inference and modeling problems in the familiar context of the HMM.

A discrete HMM has n observed states $Y_1, \ldots, Y_n$ taking on l possible values and n hidden states $X_1, \ldots, X_n$ taking on k possible values. The HMM can be characterized by the following conditional independence statements for $i = 1, \ldots, n$:

$$p(X_i|X_1, X_2, \ldots, X_{i-1}) = p(X_i|X_{i-1}),$$

$$p(Y_i|X_1, \ldots, X_i, Y_1, \ldots, Y_{i-1}) = p(Y_i|X_i).$$

We consider the homogeneous model with uniform initial distribution, where all transitions $X_i \to X_{i+1}$ are given by the same $k \times k$ matrix $S = (s_{ij})$ and all transitions $X_i \to Y_i$ are given by the same $k \times l$ matrix $T = (t_{ij})$. Throughout our discussion, we disregard for simplicity the usual probabilistic hypothesis that S and T are nonnegative and that all row sums are 1.

Proposition 1. *The HMM is the image of a map* f: $\mathbf{R}^d \to \mathbf{R}^{l^n}$, *where* $d = k(k+1)$ *and each coordinate of* f *is a bihomogeneous polynomial of degree* $n - 1$ *in* S *and degree* n *in* T.

Problem 3 is to compute the fibers of the map f. In statistics, this computation is called *parameter identification*. We use the term *coordinate polynomials* for the polynomials f_σ that are coordinates of the map f.

Our running example in this section is the case $n = 3$ with binary random variables ($k = l = 2$). The graph of this model is given in Fig. 1. The shaded nodes are the observed random variables.

Here, the parameter space is $\mathbf{R}^8$ with the coordinates s_{00}, s_{01}, s_{10}, s_{11}, t_{00}, t_{01}, t_{10}, and t_{11}, and it maps to $\mathbf{R}^8$ with the coordinates p_{000}, p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, and p_{111}. The map f: $\mathbf{R}^8 \to \mathbf{R}^8$ is given by the following:

$$\begin{aligned} f_{\sigma_1\sigma_2\sigma_3} = {} & s_{00}s_{00}t_{0\sigma_1}t_{0\sigma_2}t_{0\sigma_3} + s_{00}s_{01}t_{0\sigma_1}t_{0\sigma_2}t_{0\sigma_3} + s_{01}s_{10}t_{0\sigma_1}t_{1\sigma_2}t_{0\sigma_3} \\ & + s_{01}s_{11}t_{0\sigma_1}t_{1\sigma_2}t_{1\sigma_3} + s_{10}s_{00}t_{1\sigma_1}t_{0\sigma_2}t_{0\sigma_3} + s_{10}s_{01}t_{1\sigma_1}t_{0\sigma_2}t_{1\sigma_3} \\ & + s_{11}s_{10}t_{1\sigma_1}t_{1\sigma_2}t_{0\sigma_3} + s_{11}s_{11}t_{1\sigma_1}t_{1\sigma_2}t_{1\sigma_3}. \end{aligned}$$

The HMM (i.e., the image of f) is the zero set of the following quartic polynomial:

$$\begin{aligned} & p_{011}^2 p_{100}^2 - p_{001}^2 p_{110}^2 + p_{000}p_{011}p_{101}^2 - p_{000}p_{101}^2 p_{110} \\ & + p_{000}p_{011}p_{110}^2 \quad - p_{001}p_{010}^2 p_{111} + p_{001}^2 p_{100}p_{111} + p_{010}^2 p_{100}p_{111} \\ & - p_{001}p_{100}^2 p_{111} - p_{000}p_{011}^2 p_{110} - p_{001}p_{011}p_{100}p_{101} \\ & - p_{010}p_{011}p_{100}p_{101} + p_{001}p_{010}p_{011}p_{110} - p_{010}p_{011}p_{100}p_{110} \\ & + p_{001}p_{010}p_{101}p_{110} + p_{001}p_{100}p_{101}p_{110} + p_{000}p_{010}p_{011}p_{111} \\ & - p_{000}p_{011}p_{100}p_{111} - p_{000}p_{001}p_{101}p_{111} + p_{000}p_{100}p_{101}p_{111} \\ & + p_{000}p_{001}p_{110}p_{111} - p_{000}p_{010}p_{110}p_{111}. \end{aligned}$$

This polynomial was found by a *Gröbner basis* computation. See the discussion on *implicitization* in section 3 of ref. 11.

In general, the polynomial functions on $\mathbf{R}^{l^n}$ that vanish on the image of f are called the *invariants of the model*. They form a prime ideal I_f. In our example, I_f is generated by the quartic polynomial given above. Problem 5 is to compute generators of the ideal I_f. When l^n and d are small, this can be done by using Gröbner bases, and in some cases it is possible to characterize I_f based on the structure of the model (see, for example, *Conjecture 13*), but in general, problem 5 is hard and the ideal I_f may remain unknown.

Here, tropical geometry comes in. The *tropicalization* of our map f is the map g: $\mathbf{R}^d \to \mathbf{R}^{l^n}$ defined by replacing products by sums and sums by minima in the formula for f. In our example ($n = 3, k = l = 2$), the tropicalization is the piecewise-linear map g: $\mathbf{R}^8 \to \mathbf{R}^8$, $(U, V) \mapsto$ with the following:

$$\begin{aligned} \delta_{\sigma_1\sigma_2\sigma_3} = \min\{ & u_{h_1h_2} + u_{h_2h_3} + v_{h_1\sigma_1} + v_{h_2\sigma_2} + v_{h_3\sigma_3}: \\ & (h_1, h_2, h_3) \in \{0,1\}^3\}. \qquad \textbf{[1]} \end{aligned}$$

This minimum is attained by the most likely hidden data $(\hat{h}_1, \hat{h}_2, \hat{h}_3)$, given the observations $(\sigma_1, \sigma_2, \sigma_3)$ and given the parameters $u_{..} = -\log(s_{..})$ and $v_{..} = -\log(t_{..})$. The sequence $(\hat{h}_1, \hat{h}_2, \hat{h}_3)$ is known as the *Viterbi sequence* in the HMM literature (12). It solves problem 2 in section 1.

The key observation, which we discuss in more detail in section 4, is that the set of parameters (U, V) that selects the Viterbi sequence $(\hat{h}_1, \hat{h}_2, \hat{h}_3)$ is the normal cone at a vertex of the Newton polytope of the polynomial $f_{\sigma_1\sigma_2\sigma_3}$. This polytope is four-dimensional, it has eight vertices, and its normal fan represents the solution to problem 4 in section 1 when $\sigma = \sigma_1\sigma_2\sigma_3$ is fixed.

We can also consider an extension of problem 4 in which $\sigma = \sigma_1\sigma_2\sigma_3$ ranges over all possible observations. The solution is given by the Newton polytope of the map f. In our example, it is a five-dimensional polytope with 398 vertices, 1,136 edges, 1,150 two-faces, 478 three-faces, and 68 facets, namely, the Minkowski sum of eight copies of the earlier four-dimensional polytope for $(\sigma_1, \sigma_2, \sigma_3) \in \{0, 1\}^3$. For a concrete numerical example, fix the parameters $U^* = \left(\begin{smallmatrix} 6 & 5 \\ 8 & 1 \end{smallmatrix}\right)$ and $V^* = \left(\begin{smallmatrix} 0 & 8 \\ 8 & 8 \end{smallmatrix}\right)$. We find the following:

If the observed string at $Y_1Y_2Y_3$ is

$$\sigma_1\sigma_2\sigma_3 = 000\ 001\ 010\ 011\ 100\ 101\ 110\ 111,$$

then the Viterbi sequence at $X_1X_2X_3$ is

$$\hat{h}_1\hat{h}_2\hat{h}_3 = 000\ 001\ 000\ 011\ 000\ 111\ 110\ 111.$$

The set of all parameters (U, V) leading to the same conclusions as (U^*, V^*) is the cone defined by the following:

$$u_{01} - u_{00} + v_{11} - v_{01} \leq 0, u_{10} - u_{11} + v_{00} - v_{10} \leq 0,$$

$$u_{00} + v_{01} - u_{10} - v_{11} \leq 0,$$

$$2u_{00} + v_{01} - u_{01} - u_{10} - v_{11} \leq 0,$$

$$2u_{11} + v_{10} + v_{11} - u_{00} - u_{01} - v_{00} - v_{01} \leq 0.$$

Our solution to the parametric inference problem with respect to all observations simultaneously consists of 398 such cones. The *tropical HMM* is the union of the images of these cones under the piecewise-linear map g: $(U, V) \mapsto \delta$. This image is a piecewise-linear set of dimension 7. The cone that contains the chosen parameters (U^*, V^*) is mapped to a seven-dimensional cone in the tropical HMM (it spans the hyperplane $\delta_{010} = \delta_{100}$), but most of the other 397 cones are mapped to lower-dimensional cones by the map g. The question of how the number 398 grows as the length n increases is addressed in *Corollary 10*.

3. Positivity and Morphisms in Tropical Geometry

We have shown that a graphical model is the image of a polynomial map f from the space of parameters to the space of joint probability distributions on the observed random variables. Furthermore, we have shown that the tropicalization of f arises naturally in solving problem 4. In this section, we study the geometry of tropicalization in the more general setting where f: $\mathbf{R}^d \to \mathbf{R}^m$ is an arbitrary polynomial map. In statistical applications, each coordinate f_σ of the map f is usually a polynomial with positive coefficients. If this condition holds, then the polynomial map f is called *positive*. We consider f to be *surjectively positive* if, in addition, f maps the positive orthant surjectively onto the positive points in the image, in symbols,

$$f(\mathbf{R}^d_{>0}) = \text{image}(f) \cap \mathbf{R}^m_{>0}. \quad \mathbf{[2]}$$

The set of all polynomial functions that vanish on the image of f is a prime ideal I_f in the polynomial ring $\mathbf{R}[p_1, \ldots, p_m]$. The closure of the image of f is the variety of the prime ideal I_f.

In tropical geometry, we replace the variety of I_f by a piecewise-linear set as follows. The *tropical variety* $\mathcal{T}(I_f)$ is the set of all weight vectors $w \in \mathbf{R}^m$ such that the initial ideal $\text{in}_w(I_f)$ contains no monomial (4, 13). By following ref. 14, we define the *positive tropical variety* $\mathcal{T}^+(I_f)$ as the set of all weight vectors $w \in \mathbf{R}^m$ such that the initial ideal $\text{in}_w(I_f)$ contains no polynomial with only positive coefficients. The tropical variety $\mathcal{T}(I_f)$ is a *polyhedral fan* in $\mathbf{R}^m$, and $\mathcal{T}^+(I_f)$ is a *polyhedral subcomplex* of $\mathcal{T}(I_f)$. This observation means that $\mathcal{T}(I_f)$ is a finite union of closed convex polyhedral cones that fit together nicely, and $\mathcal{T}^+(I_f)$ is the union of a subset of these cones. The *tropicalization* of the polynomial map f is the piecewise-linear map g: $\mathbf{R}^d \to \mathbf{R}^m$ defined by replacing products by sums and sums by minima in the evaluation of f. We consider g to be a *tropical morphism*. Examples of tropical morphisms appear in Eqs. **1**, **3**, **4**, **9**, and **10**.

The following theorem describes the geometry of this situation. We define the *Newton polytope* of a polynomial map f: $\mathbf{R}^d \to \mathbf{R}^m$ as the Minkowski sum in $\mathbf{R}^d$ of the Newton polytopes of its coordinates $f_1, \ldots, f_m$. For basic information on Newton polytopes and their normal fans, see section 1 of ref. 13.

Theorem 2. *The tropical morphism g is linear on each cone in the normal fan of the Newton polytope of f. Its image is a fan contained in* $\mathcal{T}(I_f)$. *If f is positive, then image(g) is a subset of* $\mathcal{T}^+(I_f)$, *but it is generally not a polyhedral subcomplex. If f is surjectively positive, then image(g)* $= \mathcal{T}^+(I_f)$.

Proof: Let P_i denote the Newton polytope of the polynomial $f_i = f_i(s_1, \ldots, s_d)$. By definition, P_i is the convex hull in $\mathbf{R}^d$ of all nonnegative lattice points $a = (a_1, \ldots, a_d) \in \mathbf{N}^d$ such that the monomial $s_1^{a_1} \cdots s_d^{a_d}$ appears with a nonzero coefficient in f_i. The piecewise-linear concave function g_i is the *support function* of the polytope P_i. Thus, $g_i(w)$ is the minimum value attained on P_i by the linear functional $a \mapsto w{\cdot}a$. In particular, the function g_i: $\mathbf{R}^d \to \mathbf{R}$ is linear on each cone in the normal fan of P_i.

The Newton polytope of the map f is the Minkowski sum $P_1 + \cdots + P_m = \{a_1 + \cdots + a_m : a_i \in P_i\}$. The normal fan of $P_1 + \cdots + P_m$ is the common refinement of the normal fans of $P_1, \ldots, P_m$. This observation shows that the function $f = (f_1, \ldots, f_m)$: $\mathbf{R}^d \to \mathbf{R}^d$ is linear on each cone of the normal fan of the Newton polytope of f. Because g is continuous, the image of g is a closed polyhedral fan in $\mathbf{R}^m$.

Consider any vector $w \in \mathbf{R}^d$. We must show that $g(w)$ lies in $\mathcal{T}(I_f)$, and if f is positive, then $g(w)$ lies in $\mathcal{T}^+(I_f)$. Let ϕ be any polynomial in the ideal I_f. If we substitute $p_1 = f_1, \ldots, p_m = f_m$ into $\phi = \phi(p_1, \ldots, p_m)$, then the result is zero. Consequently, if we substitute the initial forms $p_1 = \text{in}_w(f_1), \ldots, p_m = \text{in}_w(f_m)$ into the initial form $\text{in}_{g(w)}(\phi)$, then the result is zero (see equation 11.2 in ref. 13, p. 100). This fact implies that $\text{in}_{g(w)}(\phi)$ is not a monomial. Moreover, if f is positive, then ϕ must have two terms whose coefficients have opposite signs.

The following example shows that image (g) need not be a subcomplex of $\mathcal{T}^+(I_f)$. If f is assumed to be surjectively positive, then it follows from proposition 2.5 in ref. 14 that image$(g) = \mathcal{T}^+(I_f)$.

Example 3: Let $d = 3$ and $m = 4$, and consider the linear map

$$f\colon \mathbf{R}^3 \to \mathbf{R}^4,\ (s_1, s_2, s_3)$$
$$\mapsto (s_1 + s_2 + s_3,\ s_1 + 2s_2 + s_3,\ s_2 + s_3,\ s_3).$$

Then, I_f is the principal ideal generated by the linear form $p_1 - p_2 + p_3 - p_4$, and $\mathcal{T}(I_f)$ is essentially the normal fan of a tetrahedron. We identify $\mathcal{T}(I_f)$ with the complete graph K_4. The six edges of K_4 are labeled with six monomial-free initial ideals of I_f, namely, $\langle p_1 + p_3\rangle$, $\langle -p_2 - p_4\rangle$, $\langle p_1 - p_2\rangle$, $\langle p_1 - p_4\rangle$, $\langle -p_2 + p_3\rangle$, $\langle p_3 - p_4\rangle$. The first two of these six initial ideals contain a polynomial with positive coefficients. Hence, the positive tropical variety $\mathcal{T}^+(I_f)$ is the four-cycle in K_4 formed by the remaining four edges.

The tropicalization of the linear map f is the tropical morphism:

$$g\colon \mathbf{R}^3 \to \mathbf{R}^4,\ (u_1, u_2, u_3) \quad \mathbf{[3]}$$
$$\mapsto [\min(u_1, u_2, u_3),\ \min(u_1, u_2, u_3),\ \min(u_2, u_3),\ u_3].$$

The image of g is the set of all vectors (a, a, b, c) with $a \leq b \leq c$. Each vector (a, a, b, c) with $a < b < c$ has the initial ideal $\langle p_1 - p_2\rangle$, so it lies on a particular edge of K_4. But the same edge also accounts for all vectors (a, a, b, c) with $a < c < b$, none of which is in the image of g. Thus, image(g) is a closed segment that covers only half of the edge of K_4 indexed by $\langle p_1 - p_2\rangle$.

In the rest of this section, we examine *Theorem 2* for a small but important graphical model, namely, the *naive Bayes model with two features* (ref. 1, section 7). There are two observed random variables Y_1 and Y_2 that depend on one hidden binary random variable X. The two observed variables take k and l possible values, respectively. The parameterization f of this model is the map f: $\mathbf{R}^{2(k+l)} \mapsto \mathbf{R}^{kl}$ given by $p_{ij} = s_{i0}t_{0j} + s_{i1}t_{1j}$. Thus, the model consists of all $k \times l$ matrices $P = (p_{ij})$ of the form $P = S{\cdot}T$, where S is a $k \times 2$ matrix and T is a $2 \times l$ matrix (i.e., the model consists of precisely the $k \times l$ matrices of rank $\leq$2).

Proposition 4. *The parameterization f of the naive Bayes model with two features is surjectively positive. The ideal* I_f *is generated by the* 3×3 *subdeterminants of the* $k \times l$ *matrix* $P = (p_{ij})$.

Proof: The map f being positive means that, if P is any positive matrix of rank 2, then S and T can be chosen to be positive. This

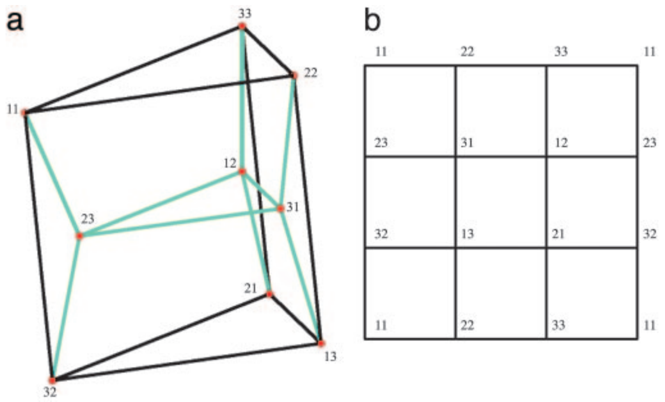

Fig. 2. The tropical variety and positive-tropical variety of the 3 × 3 determinant.

is a known result in linear algebra (e.g, see ref. 15). The same statement is false for rank ≥3 (i.e., the parameterization of the naive Bayes model with three or more features is not surjectively positive). A well known result in commutative algebra states that the $(r + 1) \times (r + 1)$ minors of a $k \times l$ matrix generate a prime ideal. The variety of this ideal is the set of $k \times l$ matrices of rank $\leq r$. This is our ideal I_f for $r = 2$.

The objects of *Theorem 2* have been studied (3, 16). The tropical variety $\mathcal{T}(I_f)$ is the set of $k \times l$ matrices of *tropical rank* ≤ 2, and the tropical variety $\mathcal{T}^+(I_f)$ = image(g) is the set of $k \times l$ matrices of *Barvinok rank* ≤ 2. Develin (16) determined the combinatorics and topology of these spaces when $\min(k, l) = 3$. He showed that $\mathcal{T}(I_f)$ is shellable but that $\mathcal{T}^+(I_f)$ can have torsion in its integral homology groups.

The Newton polytope of the map f is an interesting combinatorial object, namely, it is the $(kl - k - l + 2)$-dimensional zonotope associated with the complete bipartite graph $K_{k,l}$. The Newton polytope of each coordinate f_{ij} is a line segment, and the zonotope is their Minkowski sum. The normal fan is the hyperplane arrangement $\{u_{i0} - u_{i1} = v_{1j} - v_{0j}\}$. Its maximal cones correspond to the acyclic orientations of the complete bipartite graph $K_{k,l}$. West (17) showed that the number of facets of such a cone can be any integer between $k + l - 1$ and kl. The total number of cones equals $\sum_{i=1}^{k} S(k, i)(-1)^{l+i} i!(i + 1)^l$, where $S(k, i)$ is the Stirling number of the second kind. Here, the tropical morphism g is given by the following:

$$g_{ij} = \min(u_{i0} + v_{0j},\, u_{i1} + v_{1j}). \qquad [4]$$

The map g: $\mathbf{R}^{2(k+l)} \mapsto \mathbf{R}^{kl}$ is piecewise linear with respect to the hyperplane arrangement.

Example 5: Let $k = l = 3$, so the two observed random variables are ternary. The prime ideal is the following:

$$I_f = \langle p_{11}p_{22}p_{33} - p_{11}p_{23}p_{32} - p_{12}p_{21}p_{33} + p_{12}p_{23}p_{31} + p_{13}p_{21}p_{32} - p_{13}p_{22}p_{31} \rangle.$$

The tropical variety $\mathcal{T}(I_f)$ is the fan over a two-dimensional polyhedral complex consisting of six triangles and nine quadrangles. This complex is the 2-skeleton of the product of two triangles, labeled as in Fig. 2*a*. This complex is shellable. The positive tropical variety $\mathcal{T}^+(I_f)$ is the subcomplex consisting of the nine quadrangles shown in Fig. 2*b*. Note that $\mathcal{T}^+(I_f)$ is a torus.

The Newton polytope of f is a five-dimensional zonotope with 230 vertices, one for each acyclic orientation of the complete bipartite graph $K_{3,3}$. The map g is linear on each of the 230 cones in the corresponding hyperplane arrangement, but it is rank-deficient on 68 of the cones. The remaining $162 = 18 \times 9$ cones are mapped onto the nine quadrangles of the torus $\mathcal{T}^+(I_f)$. Thus, the general fiber of g involves 18 cones. Of these 18 cones, 8 cones have five facets, 8 cones have six facets, and 2 cones have eight facets.

4. Newton Polytopes of Graphical Models and their Complexity

Consider a graphical model with E edges and n observed random variables $Y_1, \ldots, Y_n$, each taking l values. Such a model is given by a positive polynomial map f: $\mathbf{R}^d \to \mathbf{R}^{l^n}$. Each coordinate f_σ of f is a polynomial of degree e in the model parameters $s_1, \ldots, s_d$. In this section, we discuss the statistical meaning and the computational complexity of the mathematical objects introduced in section 3.

We write $u_i = -\log(s_i)$ for the negative logarithms of the model parameters. Consider any of the l^n possible observations σ. The quantity $f_\sigma(s_1, \ldots, s_d)$ is the probability of making this particular observation [i.e., it is Prob($\mathbf{Y} = \sigma$)]. The quantity $g_\sigma(u_1, \ldots, u_d)$ is the negative logarithm of the conditional probability Prob($\mathbf{X} = \hat{\mathbf{h}} | \mathbf{Y} = \sigma$), where $\mathbf{h}$ maximizes Prob($\mathbf{X} = \mathbf{h} | \mathbf{Y} = \sigma$) for the parameters $(s_1, \ldots, s_d)$. Clearly, the function g_σ: $\mathbf{R}^d \to \mathbf{R}$ is piecewise linear and concave on the logarithmic parameter space.

The domains of linearity of the function g_σ are the cones in the normal fan of the Newton polytope of f_σ. Each maximal cone $\mathcal{C}$ is indexed by the hidden data $\hat{\mathbf{h}}$ that maximizes Prob($\mathbf{X} = \mathbf{h} | \mathbf{Y} = \sigma$) for any of the parameters $(u_1, \ldots, u_d) \in \mathcal{C}$. The hidden data $\hat{\mathbf{h}}$ that arise in this manner, for some choice of logarithmic parameters u, are called the possible *explanations* of the observation σ. For example, for the HMM described in section 2, the explanations are the Viterbi sequences.

We vary the observations as follows. Each logarithmic parameter vector $\mathbf{u}$ defines an *inference function* $\sigma \mapsto \hat{\mathbf{h}}$ from the set of observations to the set of explanations. For the HMM, each inference function $\{1, \ldots, l\}^n \to \{1, \ldots, k\}^n$ takes an observed sequence σ to the corresponding Viterbi sequence $\hat{\mathbf{h}}$. There are $(k^n)^{l^n} = k^{nl^n}$ such functions, but most of these are not inference functions. For example, consider the binary HMM of length three. There are $8^8 = 16{,}777{,}216$ Boolean functions $\{0, 1\}^3 \to \{0, 1\}^3$, but as we show at the end of section 2, only 398 of these are inference functions for the HMM.

Proposition 6. *The inference functions* $\sigma \mapsto \hat{\mathbf{h}}$ *of a graphical model f are in bijection with the vertices of the Newton polytope of the map f. The explanations* $\hat{\mathbf{h}}$ *for a fixed observation* σ *in a graphical model are in bijection with the vertices of the Newton polytope of the polynomial* f_σ.

In applications of graphical models, the number d of parameters and the number l of values of the observed random variables are small and fixed, but the number n of observed random variables is large. Recall that the model is the image of the map f: $\mathbf{R}^d \to \mathbf{R}^{l^n}$. Hence, the dimension of the model remains fixed, but the dimension of its ambient space grows exponentially in n. Therefore, it is algorithmically infeasible to compute the full tropical variety $\mathcal{T}(I_f)$. However, we can efficiently compute the Newton polytopes of the f_σ, or even the Newton polytope of f. This insight allows us to glean information about the tropical variety from the domains of linearity of its "coordinate functions" g_σ.

Our next goal is to derive an upper bound on the number of vertices of the Newton polytopes.

Theorem 7. *Consider graphical models f whose number of parameters d is fixed and whose number n of observed random variables and number of edges E varies. (Typically, E is a linear function of n.) Then, the number of vertices of the Newton polytope* $NP(f_\sigma)$ *of* f_σ *is bounded above by the following:*

$$\text{No. of vertices } (NP(f_\sigma)) \leq \text{constant}\cdot E^{d(d-1)/(d+1)} \leq \text{constant}\cdot E^{d-1}.$$

For many important families of graphical models, the number E of edges is bounded by a linear function in terms of the number n of observed nodes, and in these cases, we can replace E by n. Hence, for any given observation σ, the number of explanations grows polynomially in n. For example, in the HMM described in section 2, we have $E = 2n - 1$, and a similar relationship holds in the tree model of section 5.

Corollary 8. *For any fixed observation in the homogeneous HMM, the number of explanations is at most $C_{k,l}{\cdot}n^{k(k+l)}$. If all random variables are binary, then the upper bound $C{\cdot}n^{10/3}$ holds.*

The proofs of *Theorem 7* and *Corollary 8* are derived from the following classical result on lattice polytopes by Andrews (18). The necessary observation is that the Newton polytope of f_σ is contained in the cube $[0, E]^d$, and the volume of this cube equals E^d.

Proposition 9. *For every fixed integer d, there exists a constant C_d such that the number of vertices of any lattice polytope P in $\mathbf{R}^d$ is bounded above by $C_d{\cdot}\text{volume}(P)^{(d-1)/(d+1)}$ (18).*

The Newton polytope of the map f was defined as the Minkowski sum of the l^n smaller Newton polytopes in *Theorem 7*. We infer the following to be naive bound on its number of vertices.

Corollary 10. *The number of inference functions of a graphical model is at most $l^{nC_d}E^{d-1}$; hence, this number scales at most singly exponentially in the complexity (n, E) of the graphical model.*

Consider the homogeneous HMM on binary random variables. Each inference function is a Boolean function $\{0, 1\}^n \rightarrow \{0, 1\}^n$, but not conversely. The number of all Boolean functions is 2^{n2^n}, which grows doubly exponentially in n. However, the number of inference functions is at most $2^{\text{polynomial}(n)}$.

In practical applications of graphical models, it may be infeasible to compute all (singly exponentially many) inference functions. Nonetheless, we believe that important insight can be gained by computing and classifying the Newton polytopes of graphical models f on few random variables. Such a study would be the polyhedral analog to the algebraic classification of ref. 1.

However, for a fixed observation σ, the size of the Newton polytope of f_σ grows polynomially with the size of the graphical model, and therefore, there is hope that the polytopes can be computed efficiently. Despite the fact that the Newton polytope of f_σ has polynomially many vertices in the size of the graphical model, the number of terms in f_σ grows exponentially. This is a potential problem because the computation of the Newton polytope requires the inspection of these terms. The following result states that the convex hull computations scale with the running time of the sum–product algorithm, which for many models of interest scales polynomially with the size of the graphical model.

Proposition 11 (Polytope Propagation). *The Newton polytopes of the polynomials f_σ can be computed recursively by using the decomposition of f_σ according to the sum–product algorithm.*

5. The General Markov Model on a Binary Tree

We conclude by illustrating the concepts that we have developed in the context of tree Markov models. These models are directed graphical models in which the graph is a directed tree τ with observed random variables $Y_1, \ldots, Y_n$ at the leaves. The naive Bayes model in section 3 is the special case in which $n = 2$. Each edge e has a different transition matrix $S^e = [s^e_{\mu\nu}]$. We consider the general model given by Allman and Rhodes (8), which means that the S^e are arbitrary distinct $l \times l$ matrices. In most applications, the transition matrices are from a special model family (for example, in phylogenetics, these may be Jukes–Cantor model or the Hasegawa–Kishino–Yano model). As before, we relax the hypothesis that transition probabilities are nonnegative and sum to 1. Hence, the $s^e_{\mu\nu}$ are distinct unknowns. For simplicity, we further assume that the tree τ is binary.

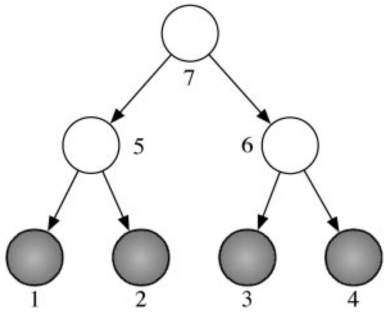

Fig. 3. A directed binary tree with $n = 4$ leaves.

Proposition 12. *The general Markov model for the binary tree τ is the image of a map f: $\mathbf{R}^{(2n-2)l^2} \rightarrow \mathbf{R}^{l^n}$, where each coordinate of f is a multilinear polynomial in the unknowns $\{(s^e_{\mu\nu}),\ e \text{ edge of } \tau\}$.*

If we denote an edge between nodes i and j by (ij), and τ' is the tree τ without the leaves, then the coordinate of the multilinear map f indexed by an observed sequence $(\sigma_1, \ldots, \sigma_n)$ can be written as follows:

$$p_{\sigma_1\cdots\sigma_n} = \sum_h \prod_{\substack{i\in\tau' \\ \text{with children } j,k}} (s^{(ij)}_{h_ih_j}{\cdot}s^{(ik)}_{h_ih_k}) \ . \qquad [5]$$

Here, h ranges over all colorations $h = (h_i)_{i\in\tau}$ of the nodes such that $h_j = \sigma_j$ for all leaves j. Our running example in this section is the binary tree in Fig. 3 with binary random variables ($l = 2$).

In this example, the coordinates of the multilinear map f: $\mathbf{R}^{24} \rightarrow \mathbf{R}^{16}$ are given by the following formula:

$$p_{\sigma_1\sigma_2\sigma_3\sigma_4} = \sum_{\{h_5,h_6,h_7\}\in\{0,1\}^3} (s^{(75)}_{h_7h_5}{\cdot}s^{(76)}_{h_7h_6}{\cdot}(s^{(51)}_{h_5\sigma_1}{\cdot}s^{(52)}_{h_5\sigma_2}){\cdot}(s^{(63)}_{h_6\sigma_3}{\cdot}s^{(64)}_{h_6\sigma_4}). \qquad [6]$$

The prime ideal I_f of polynomial invariants is generated by the 3×3 subdeterminants of the following matrix:

$$\begin{pmatrix} p_{0000} & p_{0010} & p_{0001} & p_{0011} \\ p_{0100} & p_{0110} & p_{0101} & p_{0111} \\ p_{1000} & p_{1010} & p_{1001} & p_{1011} \\ p_{1100} & p_{1110} & p_{1101} & p_{1111} \end{pmatrix} \qquad [7]$$

Thus, this particular model is the $k = l = 4$ instance of the determinantal variety in *Proposition 4*.

We generalize the determinantal presentation in this example by proposing the following explicit solution to problem 5 for arbitrary binary trees τ. Every edge of τ induces a *split* of the set of leaves $\{1, 2, \ldots, n\}$, corresponding to the two connected components of the tree obtained by removing that edge. The unrooted tree underlying τ is uniquely determined by the set of these splits.

Conjecture 13. *The ideal I_f of phylogenetic invariants of the general Markov model for any binary tree τ on binary random variables is generated by the 3×3-determinants of all two-dimensional matrices obtained by flattening the $2 \times \cdots \times 2$-table $(p_{\sigma_1\cdots\sigma_n})$ according to the splits induced by the edges of τ.*

We need to explain the meaning of the word "flattening." If (A, B) is any split of the set $\{1, \ldots, n\}$, then this term refers to the $2^{\#(A)} \times 2^{\#(B)}$ matrix whose rows and columns are indexed by the functions $A \rightarrow \{0, 1\}$ and $B \rightarrow \{0, 1\}$, respectively, and whose entries are the 2^n probabilities $p_{\sigma_1\cdots\sigma_n}$.

The sum–product algorithm is used in practice to evaluate the polynomial of Eq. **5**. Its running time is linear in n, despite the fact that the number l^{n-1} of terms in Eq. **5** grows exponentially.

This reduction in complexity is achieved by recursively grouping subsums. For example, Eq. **6** becomes the following:

$$p_{\sigma_1\sigma_2\sigma_3\sigma_4} = \sum_{v=0}^{1} (s_{v0}^{(75)} s_{0\sigma_1}^{(51)} s_{0\sigma_2}^{(52)} + s_{v1}^{(75)} s_{1\sigma_1}^{(51)} s_{1\sigma_2}^{(52)}) \cdot (s_{v0}^{(76)} s_{0\sigma_3}^{(63)} s_{0\sigma_4}^{(64)} + s_{v1}^{(76)} s_{1\sigma_3}^{(63)} s_{1\sigma_4}^{(64)}). \qquad [8]$$

Remember the following rule: Polynomials are evaluated recursively as sums of products of smaller polynomials. This is the solution to problem 1. For details on the tree case, see ref. 19.

Problem 2 is known in phylogeny as the *joint ancestral reconstruction* problem, which asks for the MAP ancestral assignments $\hat{h}_i$ given the observations $(\sigma_1, \ldots, \sigma_n)$ at the leaves. An efficient method for solving this problem is given in ref. 20. This method is nothing but the sum–product algorithm with ordinary arithmetic (+, ×) replaced by tropical arithmetic (min, +). The σ coordinate of the tropicalization g: $\mathbf{R}^{(2n-2)l^2} \rightarrow \mathbf{R}^{l^n}$ of the map in Eq. **5** is

$$\delta_{\sigma_1 \cdots \sigma_n} = \min_{h} \sum_{\substack{i \in \tau' \\ \text{with children } j,k}} (v_{h_i h_j}^{(ij)} + v_{h_i h_k}^{(ik)}). \qquad [9]$$

This expression can be evaluated efficiently by the same scheme as used previously. The rule is now the following: Piecewise-linear concave functions are evaluated recursively as minima of sums of smaller such functions. A simple example illustrating this rule is the following tropicalization of Eq. **8**:

$$\delta_{\sigma_1\sigma_2\sigma_3\sigma_4} = \min_{v \in \{0,1\}} (u_{v\sigma_1\sigma_2} + u_{v\sigma_3\sigma_4}), \qquad [10]$$

where $u_{v\sigma_1\sigma_2} = \min(v_{v0}^{(75)} + v_{0\sigma_1}^{(51)} + v_{0\sigma_2}^{(52)}, v_{v1}^{(75)} + v_{1\sigma_1}^{(51)} + v_{1\sigma_2}^{(52)})$ and similarly for $u_{v\sigma_3\sigma_4}$.

In section 4, we showed that the number of vertices of the Newton polytopes of the coordinate polynomials f_σ is critical for efficient parametric inference. That number grows polynomially in n if the number of parameters is fixed (because of *Theorem 7*), but it may grow exponentially if the number of parameters is not bounded. For the general Markov model on a tree τ, the growth will be exponential unless we restrict the number of parameters. This can be done, for example, by considering the *homogeneous tree model* as follows, where the transition matrices along all edges are identical: $s_{\mu\nu}^{e} = s_{\mu\nu}$ is independent of the edge e. By using *Theorem 7*, we obtain the following result analogous to *Corollary 8*.

Proposition 14. *The number of vertices of the Newton polytope of any coordinate f_σ in the homogeneous tree model is bounded above by n^{l^2-1} times a constant depending only on l.*

For tree models that are used in applications, such as phylogenetics, the number of parameters is likely to be reduced even further. In such cases, the parametric joint ancestral reconstruction problem can be solved efficiently by using the polytope propagation algorithm techniques given in *Proposition 11*.

6. Summary: A Statistics–Geometry Dictionary

The algebraic representation for graphical models with hidden variables leads naturally to an interpretation of a parameterized model as a point on an algebraic variety. Marginal probabilities are coordinates of points on the variety. Varieties can be tropicalized, and the statistical meaning is that the MAP probabilities (calculated with logarithms of the parameters) can be interpreted as coordinates of points on the positive part of the tropical variety. Hence, the tropical model is fundamental for understanding MAP probabilities. Although we have not addressed it in this article, the logarithms of the marginal probabilities are coordinates of points on the *amoeba* (21) of the model. Amoebas are likely to be important for understanding the geometry of maximum–likelihood estimation.

The sum–product algorithm for graphical models is an efficient method for evaluating the coordinate polynomials of a graphical model. This algorithm works in exactly the same way for classical arithmetic (+, ×) and for tropical arithmetic (min, +). The same method is used to evaluate coordinates of points on the variety and of points on the tropical variety.

An explanation for an observation σ is a vertex of the Newton polytope of f_σ. Thus, the parametric inference problem is solved by finding the normal fans of the Newton polytopes of the coordinate polynomials. For many important applications, the number of vertices of the polytopes is polynomial in the size of the graphical model. The polytope propagation algorithm, which is a geometric analog of the sum–product algorithm, finds the Newton polytopes and is efficient when the sum–product algorithm is fast and the number of vertices on the Newton polytopes is small.

In our companion article (10), we show that polytope propagation is practical and useful in the important application of biological sequence analysis. In particular, existing parametric alignment methods (22–24) can be viewed as special cases of parametric inference for a pair HMM. The computation of the Newton polytopes is also useful for Bayesian computations, where we have priors on the parameters and it is of interest to integrate over the maximal cones in the normal fan of the Newton polytope (ref. 10, section 5).

We thank Komei Fukuda, Michael Joswig, and Kristian Ranestad for their help in obtaining the computational results reported in section 2. L.P. was supported in part by National Institutes of Health Grant R01-HG02362-02. B.S. was supported by a Hewlett Packard Visiting Research Professorship 2003/2004 at the Mathematical Sciences Research Institute (MSRI) at the University of California, Berkley, and in part by National Science Foundation Grant DMS-0200729.

1. Garcia, L. D., Stillman, M. & Sturmfels, B. (2003) arXiv: math.AG/0301255.
2. Pistone, G., Riccomagno, E. & Wynn, H. P. (2001) *Algebraic Statistics: Computational Commutative Algebra in Statistics* (Chapman & Hall, Boca Raton, FL).
3. Develin, M., Santos, F. & Sturmfels, B. (2003) arXiv: math.CO/0312114.
4. Richter-Gebert, J., Sturmfels, B. & Theobald, T. (2004) in *Idempotent Mathematics and Mathematical Physics*, eds. Litvinov, G. L. & Maslov, V. P. (Am. Math. Soc., Providence, RI).
5. Jordan, M. I. & Weiss, Y. (2002) in *Handbook of Brain Theory and Neural Networks* ed. Arbib, M. (MIT Press, Cambridge, MA), 2nd Ed.
6. Kschischang, F., Frey, B. & Loeliger, H. A. (2001) *IEEE Trans. Inform. Theory* **47,** 498–519.
7. Aji, S. & McEliece, R. J. (2000) *IEEE Trans. Inform. Theory* **46,** 325–343.
8. Allman, E. & Rhodes, J. (2003) *Math. Biosci.* **186,** 113–144.
9. Cavender, J. & Felsenstein, J. (1987) *J. Classif.* **4,** 57–71.
10. Pachter, L. & Sturmfels, B. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 16138–16143.
11. Cox, D., O'Shea, D. & Little, J. (1996) *Ideals, Varieties and Algorithms* (Springer, New York).
12. Rabiner, L. R. (1989) *Proc. IEEE* **77,** 257–286.
13. Sturmfels, B. (1996) *Gröbner Bases and Convex Polytopes* (Am. Math. Soc., Providence, RI), Vol. 8.
14. Speyer, D. & Williams, L. (2003) arXiv: math.CO/0312297.
15. Cohen, J. & Rothblum, U. (1993) *Linear Algebra Appl.* **190,** 149–168.
16. Develin, M. (2004) arXiv: math.CO/0401224.
17. West, D. (1995) *Discrete Math.* **138,** 393–396.
18. Andrews, G. (1963) *Trans. Am. Math. Soc.* **106,** 270–273.
19. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998) *Biological Sequence Analysis (Probabilistic Models of Proteins and Nucleic Acids)* (Cambridge Univ. Press, Cambridge, U.K.).
20. Pupko, T., Pe'er, I., Shamir, R. & Graur, D. (2000) *Mol. Biol. Evol.* **17,** 890–896.
21. Viro, O. (2002) *Notices Am. Math. Soc.* **49,** 916–917.
22. Fernández-Baca, D., Seppäläinen, T. & Slutzki, G. (2000) in *Combinatorial Pattern Matching, Lecture Notes in Computer Science*, eds. Giancarlo, R. & Sankoff, D. (Springer, Berlin), Vol. 1,848, pp. 68–82.
23. Gusfield, D., Balasubramanian, K. & Naor, D. (1994) *Algorithmica* **12,** 312–326.
24. Waterman, M., Eggert, M. & Lander, E. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 6090–6093.