

Algebraic Geometers See Ideal Approach to Biology

By Barry A. Cipra

Doug Arnold, director of the Institute for Mathematics and its Applications, takes a broad view of applied mathematics, consistent with his conviction that mathematics shows up everywhere. Still, Arnold admits that as recently as five years ago, he could not have imagined an IMA workshop highlighting applications of algebraic geometry to evolutionary biology. Robotics? Sure. Error-correcting codes? Why not? But phylogenetic trees? Doesn't sound likely!

How time flies.

As part of its 2006–07 program on applications of algebraic geometry, the IMA had scheduled an early-March workshop on “emerging” applications. That event morphed into a workshop on algebraic biology and statistics, where more than a hundred participants learned about the roles such seeming arcana as toric varieties, projective space, polytopes, and polynomials over finite fields are beginning to play in such seeming arcana as phylogenetics, epistasis, conditional distributions, and contingency tables. Many of the talks and posters focused on the rapidly evolving role of algebraic geometry in studies of evolutionary relationships of living (and extinct) organisms.

Algebraic geometry is, at heart, the study of systems of multivariate polynomial equations and their simultaneous solution sets. Polynomiality brings with it the abstraction of algebra, with its rings and ideals. Solution sets introduce elements of geometry, along with topological notions of continuity, connectedness, closure, and so forth. Their interplay often makes it possible to find simple algebraic answers to hard geometry problems on one hand, and simple geometric answers to hard algebra problems on the other.

With pure mathematics at the helm through much of the 20th century, algebraic geometry followed a course of increasing abstract generality, with curves and surfaces giving way to schemes and sheaves. Computational aspects were kept below-decks—in part because, except for the strictly linear case, systems of polynomial equations are extremely messy to solve. But the advent of computer algebra systems and the development of algorithms implementing the theory of Gröbner bases (a nonlinear generalization of Gaussian elimination) have brought applications back to the fo'c'sle. We'll stop short of identifying mathematical equivalents of Captains Bligh and Queeg.

The prospect of doing (presumably) mistake-free calculations with huge systems of polynomials has led algebraic geometers to look toward applications. At the same time, researchers whose models spawn such systems have been encouraged to look at what theory has to offer. Phylogenetics poses some of the greatest challenges, paired with great possibilities.

The Tree(s) of Life

Broadly speaking, phylogenetics is the art—or, increasingly, the science—of inferring past relationships from present similarities. The ultimate aim is to map out the entire evolutionary “tree of life,” from its ribonucleic roots to its ephemeral buds and blossoms. For the nonce, though, re-searchers are content to sketch out a few twigs here and there. They want to know, for example, how humans, gorillas, and chimpanzees are related. (Actually this question is scientifically uncontroversial. Humans and chimps are virtually siblings, gorillas being more like kissing cousins. Even so, details of the speciations are still being explored. Humans are closer to gorillas in certain respects, much as a child can have Uncle Bob's nose or Aunt Betty's eyes.)

Gene-sequencing technology has brought a sea change to the study of phylogenetics. While comparative anatomy continues to provide valuable evolutionary clues, the gold standard now is the careful alignment of DNA, those iconic strings of As, Cs, Gs, and Ts, from different species. In extreme caricature (ignoring, for example, the subtle but, to biologists, important difference between “gene trees” and “species trees”), the tree of life sprouts a new branch whenever the offspring of some creature has DNA different from its parent's at a single site. In reality, evolution is not so straightforward. In addition to simple “point mutations,” the DNA of an offspring can experience insertions, deletions, rearrangements, and a host of other, mostly teratogenic, assaults. And as usual, sex complicates everything. The alignment of DNA is an art (and/or science) unto itself. But given an alignment, polynomials kick in.

In a pair of talks, Elizabeth Allman of the University of Alaska and Marta Casanellas of the Universitat Politècnica de Catalunya in Barcelona explained how. (In a third talk, Jaroslaw Wisniewski of the University of Warsaw gave a geometer's view of phylo-genetic trees.) The basic idea is to posit a tree, assign a random variable, taking values A, C, G, and T, to each node and a Markovian transition matrix to each edge, and then see how well the resulting probability distribution at the leaves matches up with the observed data, assuming the mutations at each site of the alignment to be independent events. Matrix multiplication along paths from the tree's root to the various leaves produces a polynomial parameterization of a d -dimensional manifold in 4^n -dimensional space, where d is the total number of variables in the parameterization and n is the number of species being compared (i.e., the number of leaves in the tree).

In most models, d grows linearly with n , so the co-dimension $4^n - d$ tends to be large. This implies a correspondingly large number of independent relationships among the polynomials describing the probability distribution—in the language of algebraic geometry, the manifold (or, more precisely, its “Zariski closure”) is an algebraic variety whose prime ideal has a large set of generators. (Just to clarify, polynomials of two types, and variables of two types, are involved: those giving the parameterization, and those relating the parameterized variables. A simple example to keep in mind is the parameterization $x = p^2 - q^2$, $y = 2pq$, $z = p^2 + q^2$ of the cone $x^2 + y^2 = z^2$.)

When push comes to shove, variables representing probabilities must be non-negative real numbers summing to 1. But most of the interest-

ing algebraic geometry ignores the probabilistic niceties, taking place instead in complex projective space. Not that algebra is necessarily complex. “It would be incredible and desirable to apply the tools of real algebraic geometry in this setting,” Allman notes. “It’s just so hard that so far no one has attempted it.”

Generators of the ideal $I(V)$ (V standing for the variety) are called “phylogenetic invariants.” The concept was introduced in the late 1980s in a pair of papers, one by Joseph Felsenstein of the University of Washington and James Cavender of the University of Colorado and Martin Marietta Aerospace in Denver, and one by James Lake of UCLA. Each polynomial in the ideal can, in principle, be interpreted as a statement about conditional probabilities; the “best” invariants have simple, biologically meaningful interpretations.

The virtue of working with polynomials from the ideal is that you can plug the observed data more or less straight in, instead of trying to find best-fit values of the parameters. A perfect fit would always produce zero; if you don’t get close to zero with all the invariants, then there are no parameter values that agree with the data for the tree posited. The goal is thus to find invariants as efficiently as possible.

Some invariants depend only on assumptions about the parameterization, such as symmetries between purines (A and G) and pyrimidines (C and T). Such invariants can be thought of as providing a reality check on the whole enterprise: If the observed distribution doesn’t satisfy these polynomial equations (approximately, of course), something is amiss.

The interesting invariants are the ones that vary from tree to tree. An “ideal” algorithm (if you’ll pardon the pun) would pinpoint these invariants, so that alternative trees could be tested against the data. That remains an elusive goal, except in the simplest of settings.

Indeed, Allman says, results of some early analyses led some to dismiss the approach. The low-hanging fruit, such as linear invariants (i.e., linear equations satisfied by the probabilities), proved to be of limited value. But recent work has begun to make larger classes of invariants available.

Group-based Models

In one major advance, Bernd Sturmfels of the University of California at Berkeley and Seth Sullivant, now at Harvard, performed a thorough analysis of phylogenetic invariants for “group-based” models. These are models in which the Markovian transition matrices can be simultaneously diagonalized by the discrete Fourier transform associated with a finite abelian group. This sounds highly abstruse and restrictive; happily, though, it accords with biologically meaningful assumptions. In particular, the so-called Jukes–Cantor model from 1969, and the Kimura models from the early 1980s, fall into this category.

The Kimura 3-parameter (K3P) model is the most general of these models. Each edge of the tree is assigned a matrix of the form

$$\begin{pmatrix} d & a & b & c \\ a & d & c & b \\ b & c & d & a \\ c & b & a & d \end{pmatrix}$$

where $d = 1 - a - b - c$. Setting b equal to c gives the K2P model. (In these models, the first two rows and columns correspond to the purines A and G, and the last two to the pyrimidines C and T. The symmetric nature of the matrix reflects the assumption that evolution, at the biochemical level, is unaware of the arrow of time.) Setting all three parameters equal gives the Jukes–Cantor model. A further, simpler model known as the Jukes–Cantor binary model has symmetric 2×2 transition matrices that care only whether a site is occupied by a purine or a pyrimidine. The finite groups lurking in the background are the virtually trivial cyclic group of order 2 and the dihedral, or Klein, group of order 4 (traditionally denoted Z_2 and $Z_2 \times Z_2$, respectively).

The change of coordinates created by the discrete Fourier transform leads to a variety parameterized by simple, single-term monomials instead of complicated, multi-term polynomials—a property first noted in the phylogenetic context in 1989, in papers by Michael Hendy and David Penny of Massey University in New Zealand. This greatly facilitates the search for invariants. Algebraic spaces of this type are known as toric varieties. (They are closely related to higher-dimensional analogues of the two-dimensional torus, or donut.) Knowledge of toric varieties and their ideals is extensive—but hardly complete.

The groundwork for studying group-based models was laid in the early 1990s by a number of researchers. Sturmfels and Sullivant have polished off the problem by finding a recursive construction of phylogenetic invariants for these models: The invariants can be stitched together from invariants for “claw” trees, also called “stars,” in which k leaves are attached directly to the root. One consequence is that the ideals for the Jukes–Cantor and Kimura models on trivalent trees, for which each of the claw trees has three leaves, are generated by invariants of degree at most 4. In particular, the Jukes–Cantor binary model for trivalent trees has purely quadratic invariants.

Sonja Petrovic and Julia Chifman, grad students at the University of Kentucky, have generalized the result on Jukes–Cantor binary models from trivalent to arbitrary claw trees; Petrovic presented their work in a poster session at the IMA workshop. This work, combined with Sturmfels and Sullivant’s results, leads to invariants for the Jukes–Cantor binary model on any tree.

The Sturmfels–Sullivant theory provides one further insight: The co-dimension of the variety gives only a lower bound on the number of phylogenetic invariants. The co-dimension, which is the difference between the dimension of the ambient space (4^n in the Kimura and Jukes–Cantor models, 2^n in the Jukes–Cantor binary model) and the number of free parameters, gives the right number of equations to describe a variety at its smooth points. The varieties that arise in practice, however, typically have singular points as well. This happens even in the simple case of a Jukes–Cantor binary model on a trivalent tree with four leaves: The co-dimension is 8, but the ideal requires 20 generators. For K3P, the co-dimension is 48, but the ideal calls for a whopping 8002 generators!

General Markov Models

What about general (non-group-based) Markov models? The elucidation of invariants is far more difficult in the absence of a Fourier trans-

form. Nevertheless, Allman and her husband, John Rhodes, who is also at the University of Alaska, have very nearly duplicated the Sturmfels–Sullivant theory for these models. The snag, surprisingly enough, is identifying the generators, in general, for the seemingly trivial three-leaf tree (see Figure 1).

More precisely, Allman and Rhodes have studied k -state general Markov models on trivalent trees. The most interesting cases phylogenetically are $k = 2$ and $k = 4$, which generalize the Jukes–Cantor binary (purine–pyrimidine) model and the Kimura (ACGT) model, respectively, but most of the theory works for any k . (Much of the theory also works for arbitrary trees, but their main result is restricted to the trivalent case.) They show that the algebraic variety for an n -leaf tree is defined by a set of polynomials obtained by “flattening” the tree along its various edges. Flattening essentially amounts to grouping the leaves on either side of the chosen edge, effectively reducing the tree to a root and two leaves (see Figure 2). Consequently, a defining set for the variety of an n -leaf tree can be constructed from polynomials defining the variety of the simple 3-leaf tree.

That variety is easy enough to identify (at least if you know a lot of algebraic geometry). It is the k -secant variety of the Segre product $P^{k-1} \times P^{k-1} \times P^{k-1}$, where P is the complex projective plane.

What’s not easy is to identify the generators of the variety’s ideal. Results have been obtained for $k = 2$ and 3, but the problem is wide open for $k = 4$ and above.

Actually, there is a second snag as well: Even if generators can be obtained for the 3-leaf tree, the Allman–Rhodes construction guarantees a defining set only for the n -leaf variety, not necessarily a complete set of generators. (Here’s a simple way to appreciate the difference: The zero set of the polynomial x^2 defines the y -axis in the Cartesian plane, but the ideal is generated by the polynomial x .) Allman and Rhodes conjecture that their construction does produce a complete set of generators, but that remains to be proved.

In the interesting $k = 4$ case, everything in the ideal is known to have degree at least 5. The quintics themselves comprise a vector space of dimension 1728. (Thomas Hagedorn of the College of New Jersey computed the dimension in 2000.) Allman and Rhodes have found an explicit construction of a spanning set for this space. It’s known that the quintics don’t generate the entire ideal—and may not even define the variety. Allman and Rhodes have shown, however, that any excess zero set the quintics may have lies in an explicitly describable set. In her talk at the IMA workshop, Allman offered a reward for anyone who can determine the ideal defining $\text{Sec}^4(P^3 \times P^3 \times P^3)$: a smoked Copper River salmon, caught by Allman herself.

The situation is much more satisfactory for $k = 2$: Allman and Rhodes have shown the ideal for the general Markov model on an n -leaf tree to be generated (not just defined) by the specified flattenings, which take the form of 3×3 minors.

Phylogenetic Inference

Hard as it is, finding invariants is only half the battle. The other, equally daunting half is using them to make phylogenetic inferences. Casanellas described work in which she and Jesus Fernández-Sánchez, a colleague at the Universitat Politècnica de Catalunya, evaluated how well invariants do at picking out the proper tree in a 4-leaf K3P model. There is only one trivalent 4-leaf tree shape, with three possible labelings (see Figure 3). In 1995, John Huelsenbeck of the University of California at San Diego compared several inference methods, including Lake’s linear invariants, testing them on a two-parameter set of simulated DNA data ranging in length from a hundred to ten thousand sites. Performance of the linear invariants was unsurprisingly poor, compared with that of an approach known as neighbor-joining, and of the more or less full-blown method of maximum likelihood (see Figure 4, rows a–c).

Casanellas and Fernández-Sánchez have revisited Huelsenbeck’s comparison, but using all 8002 generators for the 4-leaf tree. (The complete set consists of 144 binomials of degree 2, 1984 binomials of degree 3, and 5874 binomials of degree 4. They can be found in a catalog of phylogenetic invariants being amassed at <http://www.math.tamu.edu/~ljp/small-trees/>; Luis Garcia Puente of Texas A&M University maintains the site.) The full set of invariants does much better, rivaling results of maximum likelihood, especially as sequence length approaches 1000. Even for sequences of length 100, the accuracy of their method exceeds 95% in a substantial portion of Huelsenbeck’s parameter space (see Figure 4, row d).

Could more be achieved with less? Very possibly so. Casanellas reported on recent work in which a select set of only 48 invariants does equally well (see Figure 4, row e). And in a poster session at the workshop, Nicholas Eriksson of Stanford University presented work with Yuan Yao,

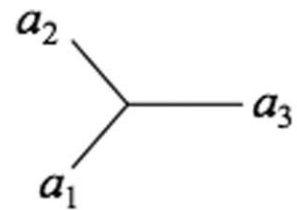


Figure 1. Help wanted. The algebraic variety for a k -state general Markov model (think $k = 4$, with the four choices being A, C, G, and T) for a trivalent tree is known, through a thick-et of definitions, theorems, and proofs, to be the k -secant variety of the Segre product $P^{k-1} \times P^{k-1} \times P^{k-1}$, where P is the complex projective plane. But what polynomials generate its ideal? Except for $k = 2$ and 3, this is an open problem.

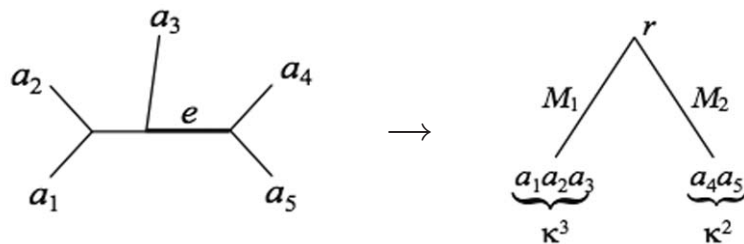


Figure 2. Tree trimming. “Flattening” a tree is one of the key techniques for constructing phylogenetic invariants. (Figure courtesy of Elizabeth Allman.)

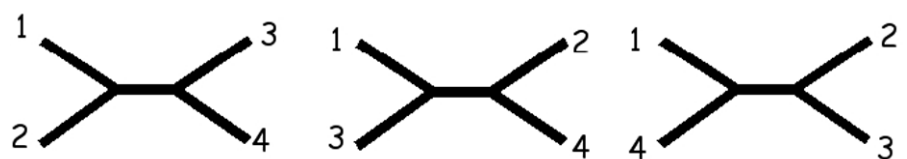


Figure 3. Take your pick. Here are three different ways to label the leaves of the only trivalent, 4-leaf tree. How well do invariants do at distinguishing them? Figure courtesy of Marta Casanellas.

also at Stanford, on a machine-learning approach that found 52 invariants (for the 4-leaf K3P model) that do even better on the benchmark parameter space. A vast array of open problems, along with purely computational hurdles, lie ahead, but algebraic geometers consider the importance of invariants well established and, like invariants themselves, not subject to change.

Barry A. Cipra is a mathematician and writer based in Northfield, Minnesota.

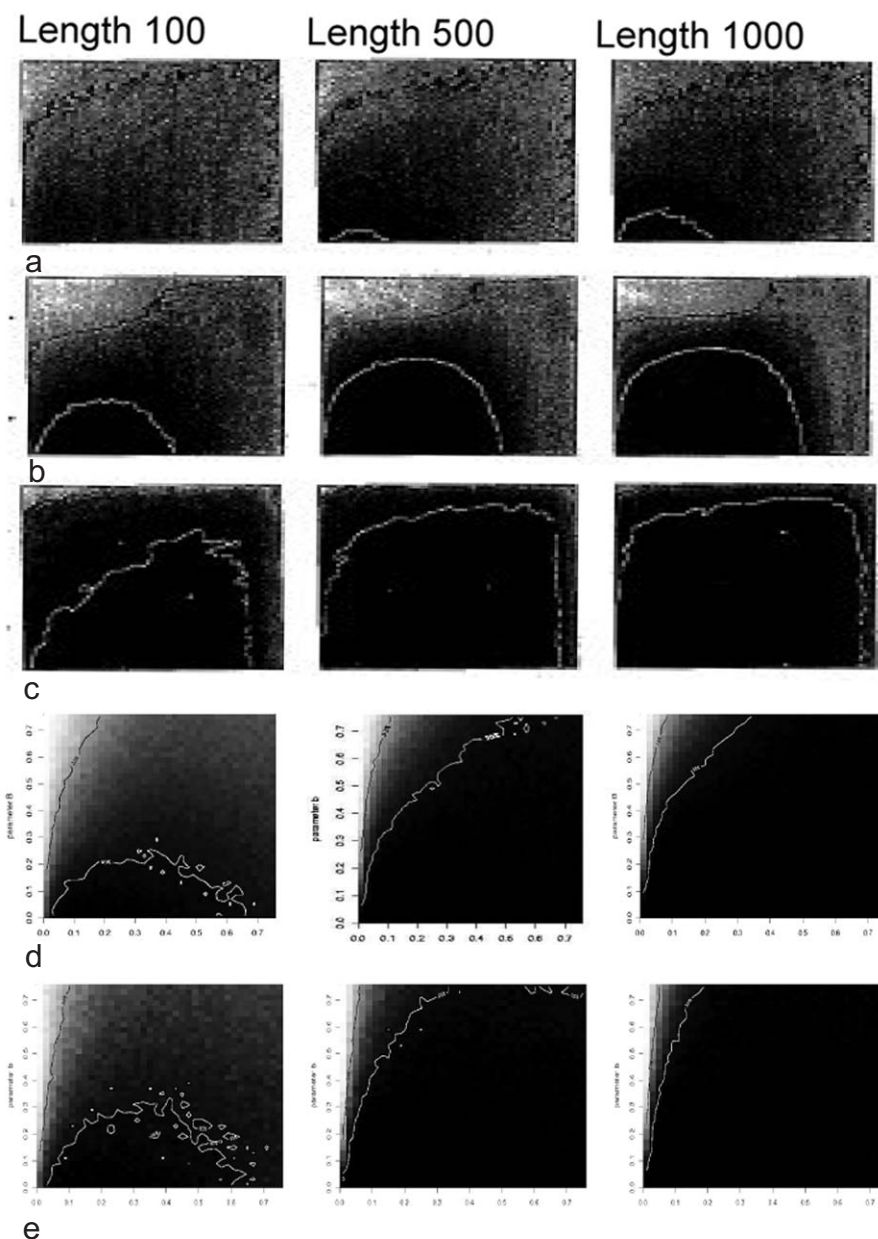


Figure 4. *Shades of gray.* A comparison by John Huelsenbeck (“Performance of phylogenetic methods in simulation,” in *Systematic Biology*, 1995, Vol. 44, No. 1, 17–48) of Lake’s linear invariants (a), a “neighbor-joining” method (b), and maximum likelihood (c) on a two-parameter family of examples for a 4-leaf tree shows the algebraic approach doing rather poorly—the black region beneath the 95% “success” isocline (jagged white line) is small, even for alignments with 1000 sites. Marta Casanellas and Jesus Fernández-Sánchez have found that using all 8002 generators does much better (d)—and a hand-picked set of 48 of these does better yet (e). (Figure courtesy of Marta Casanellas.)