CAN BIOLOGY LEAD TO NEW THEOREMS?

BERND STURMFELS

Abstract. This article argues for an affirmative answer to the question in the title. In future interactions between mathematics and biology, both fields will contribute to each other, and, in particular, research in the life sciences will inspire new theorems in "pure" mathematics. This point is illustrated by a snapshot of four recent contributions from biology to geometry, combinatorics and algebra.

Much has been written about the importance of mathematics for research in the life sciences in the 21st century. Universities are eager to start initiatives aimed at promoting the interaction between the two fields, and the federally funded mathematics institutes (AIM, IMA, IPAM, MBI, MSRI, SAMSI) are outdoing each other in offering programs and workshops at the interface of mathematics and the life sciences. The Clay Mathematics Institute has had its share of such programs. For instance, in the summer of 2005, two leading experts, Charles Peskin and Simon Levin, served as Clay Senior Scholars in the *Mathematical Biology* program at the IAS/Park City Mathematics Institute (PCMI), and in November 2005, Lior Pachter, Seth Sullivant and the author organized a workshop on *Algebraic Statistics and Computational Biology* at the Clay Mathematics Institute in Cambridge.

Yet, as these ubiquitous initiatives and programs unfold, many mathematicians remain unconvinced, and some secretly hope that this "biology fad" will simply go away soon. They have not seen any substantive impact of quantitative biology in their area of expertise, and they rightfully ask: **where are the new theorems?**

In light of these persistent doubts, some long-term observers wonder whether anything has really changed in the twenty years since Gian-Carlo Rota wrote his widely quoted sentence, *"The lack of real contact between mathematics and biology is either a tragedy, a scandal, or a challenge, it is hard to decide which"* [16, page 2]. Of course, Rota was well aware of the long history of mathematics helping biology, such as the development of population genetics by Fisher, Hardy, Wright and others in the early 1900's. Nonetheless, Rota concluded that there was no "real contact".

But, quite recently, other voices have been heard. Some scholars have begun to argue that "real contact" means being equal partners, and that meaningful intellectual contributions can, in fact, flow in both directions. This optimistic vision is expressed succinctly in the title of J.E. Cohen's article [6]: *"Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better"*.

Physics remains the gold standard for mathematicians, as there has been "real contact" and mutual respect over a considerable period of time. Historically, mathematics has made many contributions to physics, and in the last twenty years there has been a payback beyond expectations. Many of the most exciting developments

in current mathematics are a direct outgrowth of research in theoretical physics. Today's geometry and topology are unthinkable without string theory, mirror symmetry and quantum field theory. It is "obvious" that physics can lead to new theorems. Any colloquium organizer in a mathematics department who is concerned about low attendence can reliably fill the room by scheduling a leading physicist to speak. The June 2005 public lecture on *Physmatics* by Clay Senior Scholar Eric Zaslow sums up the situation as follows: *"The interplay between mathematics and physics has, in recent years, become so profound that the lines have been blurred. The two disciplines, long complementary, have begun a deep and fundamental relationship..."*.

Will biology ever be mathematics' next physics? In the future, will a theoretical biologist ever win a Fields medal? As unlikely as these possibilities seem, we do not know the answer to these questions. However, my recent interactions with computational biologists have convinced me that there is more potential in this regard than many mathematicians may be aware of. In what follows I wish to present a personal answer to the legitimate question: **where are the new theorems?**

I shall present four theorems which were inspired by biology. These theorems are in algebra, geometry and combinatorics, my own areas of expertise. I leave it to others to discuss biology-inspired results in dynamical systems and partial differential equations. Before embarking on the technical part of this article, the following disclaimer must be made: the mathematics presented below is just a tiny first step. The objects and results are certainly not as deep and important as those in Zaslow's lecture on Physmatics. But then, Rome was not built in a day.

We start our technical discussion with a contribution made by evolutionary biology to the study of metric spaces. This is part of a larger theory developed by Andreas Dress and his collaborators [2, 9, 10]. A *finite metric space* is a symmetric $n \times n$-matrix $D = (d_{ij})$ whose entries are non-negative ($d_{ij} = d_{ij} \geq 0$), zero on the diagonal ($d_{ii} = 0$), and satisfy the triangle inequalities ($d_{ik} \leq d_{ij} + d_{jk}$). Each metric space D on $\{1, 2, \ldots, n\}$ is a point in $\mathbb{R}^{\binom{n}{2}}$. The set of all such metrics is a full-dimensional convex polyhedral cone in $\mathbb{R}^{\binom{n}{2}}$, known as the *metric cone* [8].

With every point D in the metric cone one associates the convex polyhedron

$$P_D \;=\; \bigl\{\, x \in \mathbb{R}^n \,:\, x_i + x_j \;\geq\; d_{ij} \;\text{ for all } i, j \,\bigr\}.$$

If $D_1, \ldots, D_k$ are metric spaces then $D_1 + \cdots + D_k$ is a metric space as well, and

$$P_{D_1+D_2+\cdots+D_k} \;\supseteq\; P_{D_1} + P_{D_2} + \cdots + P_{D_k}.$$

If this inclusion of polyhedra is an equality then we say that the sum $D_1+D_2+\cdots+D_k$ is *coherent*. A *split* is a pair (α, β) of disjoint non-empty subsets of $\{1, \ldots, n\}$ such that $\alpha \cup \beta = \{1, \ldots, n\}$. Each split (α, β) defines a *split metric* $D^{\alpha,\beta}$ as follows:

$$D^{\alpha,\beta}_{ij} = 0 \;\text{ if }\; \{i,j\} \subseteq \alpha \text{ or } \{i,j\} \subseteq \beta, \quad \text{ and } D^{\alpha,\beta}_{ij} = 1 \;\text{ otherwise.}$$

The polyhedron $P_{D^{\alpha,\beta}}$, which represents a split metric $D^{\alpha,\beta}$, has precisely one bounded edge, and its two vertices are the zero-one incidence vectors of α and β. A metric D is called *split-prime* if it cannot be decomposed into a coherent sum of a positive multiple of a split metric and another metric. The smallest example of a split-prime metric has $n = 5$, and it is given by the distances among the nodes in the complete bipartite graph $K_{2,3}$.

Theorem 1. (Dress-Bandelt Split Decomposition [2]**)** *Every finite metric space D admits a unique coherent decomposition $D = D_1 + \cdots + D_k + D'$, where $D_1, \ldots, D_k$ are linearly independent split metrics and D' is a split-prime metric.*

This theorem is useful for evolutionary biology because it offers a polyhedral framework for phylogenetic reconstruction. Suppose we are given n taxa, for instance the genomes of n organisms, and we take D be a matrix of distances among these taxa. In typical applications, d_{ij} would be the Jukes-Cantor distance [21, §4.4] derived from a pairwise alignment of genome i and genome j. Then we consider the polyhedral complex $\mathrm{Bd}(P_D)$ whose cells are the bounded faces of the polyhedron P_D. This is a contractible complex known as the *tight span* [9] of the metric space D. The metric D is a *tree metric* if and only if the tight span $\mathrm{Bd}(P_D)$ is one-dimensional, and, in this case, the one-dimensional contractible complex $\mathrm{Bd}(P_D)$ is precisely the *phylogenetic tree* which represents the metric D.

The space of phylogenetic trees on n taxa was introduced by Billera, Holmes and Vogtmann [4]. Since every tree metric uniquely determines its tree, this space is a subset of the metric cone. It can be characterized as follows:

Corollary. *The space of trees of* [4] *equals the following subset of the metric cone:*

$$\mathrm{Trees}_n \quad = \quad \bigl\{ D \in \mathbb{R}^{\binom{n}{2}} \,:\, D \textit{ is a metric} \;\; \textit{and} \;\; \textit{dim}\,\mathrm{Bd}(P_D) \,\leq\, 1 \bigr\}.$$

If the metric D arises from real data then it is unlikely to lie exactly in the space of trees. Standard methods used by biologists, such as the neighbor joining algorithm, compute a suitable projection of D onto Trees_n. From a mathematical point of view, however, it is desirable to replace the concept of a tree by a higher-dimensional object that faithfully represents the data. The tight span $\mathrm{Bd}(P_D)$ is the universal object of this kind. It can be computed using the software `POLYMAKE`. Figure 2 shows the tight span of a metric on six taxa. This metric was derived from an alignment of DNA sequences of six bees. For details and an introduction to `POLYMAKE` we refer to [14]. We note that, for larger data sets, the tight span is often too big. This is where Theorem 1 enters the scene: what one does is remove the *splits residue* D' from the data D. The remaining split-decomposable metric $D_1 + \cdots + D_k$ can be computed efficiently with the software `SPLITSTREE` due to Huson and Bryant [15]. It is represented by a *phylogenetic network*.

Andreas Dress now serves as director of the Institute for Computational Biology in Shanghai (`www.icb.ac.cn`), a joint Chinese-German venture. He presented his theory at the November 2005 workshop at the Clay Mathematics Institute in Cambridge. In his invited lecture at the 1998 ICM in Zürich, Dress suggested that the *"the tree of life is an affine building"* [10]. Affine buildings are highly symmetric infinite simplicial complexes which play an important role in several areas of mathematics, including group theory, representation theory, topology and harmonic analysis.

The insight that phylogenetic trees, and possible higher-dimensional generalizations thereof, are intimately related to affine buildings is an important one. The author of this article agrees enthusiastically with Dress' point of view, as it is consistent with recent advances at the interface of phylogenetics and tropical geometry. An interpretation of tree space as a Grassmannian in tropical algebraic geometry

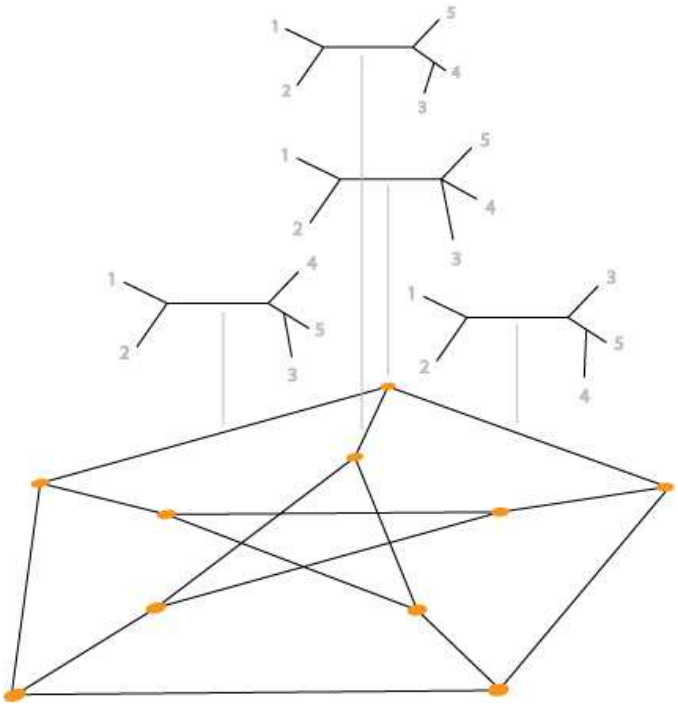

FIGURE 1. The space of phylogenetic trees on five taxa is a seven-dimensional polyhedral fan inside the ten-dimensional metric cone. It has the combinatorial structure of the Petersen graph, depicted here. The fan Trees_5 consists of 15 maximal cones, one for each edge of the graph, which represent the trivalent trees. They meet along 10 six-dimensional cones, one for each vertex of the graph.

was given in [24]: Figure 1 really depicts a Grassmannian together with its tautological vector bundle. It is within this circle of ideas that the next theorem was found, three years ago, by Lior Pachter and Clay Research Fellow David Speyer [20].

Let T be a phylogenetic tree with leaves labeled by $[n] = \{1, 2, \ldots, n\}$, and with a non-negative length associated to each edge of T. Then we define a real-valued function $\delta^{T,m}$ on the m-element subsets I of $[n]$ as follows: the number $\delta^{T,m}(I)$ is the sum of the lengths of all edges in the subtree spanned by I. For $m = 2$ we recover the tree metric $D_T = \delta^{T,2}$. We call $\delta^{T,m} : \binom{[n]}{m} \to \mathbb{R}$ the *subtree weight function*.

Theorem 2. (Pachter-Speyer Reconstruction from Subtree Weights [20]**)** *Suppose that $n \geq 2m-1$. Every phylogenetic tree on n taxa is uniquely determined by its subtree weight function. More precisely, $\delta^{T,m}$ determines the tree metric $\delta^{T,2}$.*

The punchline of this theorem is a statistical one. The aim of replacing $m = 2$ by larger values of m is that $\delta^{T,m}$ can be estimated from data in a more reliable manner. Practical advantages of this method were shown in [19].

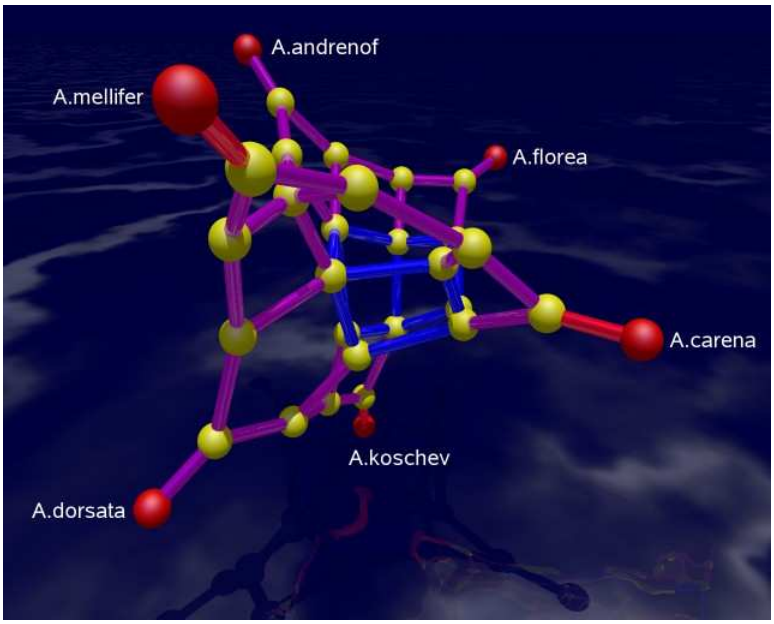

FIGURE 2. The tight span of a six-point metric space derived from aligned DNA sequences of six species of bees. We thank Michael Joswig and Thilo Schröder for drawing this diagram and allowing us to include it. See [14] for a detailed description.

Phylogenetics has spawned several different research directions in current mathematics, especially in combinatorics and probability. For more information, we recommend the book by Semple and Steel [23], and the special semester on Phylogenetics which will take place in Fall 2007 at the Newton Institute in Cambridge, England.

Algebraists, geometers and topologists may also enjoy a glimpse of *phylogenetic algebraic geometry* [13]. Here the idea is that statistical models of biological sequence evolution can be interpreted as algebraic varieties in spaces of tensors. This approach has led to a range of recent developments which are of interest to algebraists; see [1, 18, 25] and the references given there. As an illustration, we present a recent theorem due to Buczynska and Wisniewski [5]. The abstract of their preprint leaves no doubt that this is an unusual paper as far as mathematical biology goes: *"We investigate projective varieties which are geometric models of binary symmetric phylogenetic 3-valent trees. We prove that these varieties have Gorenstein terminal singularities (with small resolution) and they are Fano varieties of index 4...."*.

The varieties studied here are all embedded in the projective space $\mathbb{P}^{2^{n-1}-1} = \mathbb{P}(\mathbb{C}^2 \otimes \mathbb{C}^2 \otimes \cdots \otimes \mathbb{C}^2)$ whose coordinates x_I are indexed by subsets I of $\{1, \ldots, n\}$ whose cardinality $|I|$ is even. We fix a trivalent tree T whose leaves are labeled by $1, \ldots, n$. Each of the $2n-3$ edges e of the tree T is identified with a projective line $\mathbb{P}^1$ with homogeneous coordinates $(u_e : v_e)$. For any even subset I of the leaves of T there exists a unique set $\text{Paths}(I)$ of disjoint paths, consisting of edges of T, whose end points are the leaves in I. This observation gives rise to a birational

morphism

$$\phi_T : (\mathbb{P}^1)^{2n-3} \to \mathbb{P}^{2^{n-1}-1} \quad \text{defined by} \quad x_I = \prod_{e \in \text{Paths}(I)} u_e \cdot \prod_{e \notin \text{Paths}(I)} v_e.$$

The closure of the image of ϕ_T is a projective toric variety which we denote by X_T.

Theorem 3. (Buczynska-Wisniewski Flat Family of Trees [5]**)** *All toric varieties X_T are the same connected component of the Hilbert scheme of projective schemes, as T ranges over all combinatorial types of trivalent trees with $n+1$ leaves. Combinatorially, this means that the convex polytopes associated with these toric varieties all share the same Ehrhart polynomial (a formula for this Ehrhart polynomial is given in* [5, §3.4]*).*

Earlier work with Seth Sullivant [25] had shown that the homogeneous prime ideal of X_T has a Gröbner basis consisting of quadrics. These quadrics are the 2×2-minors of a collection of matrices, two for each edge e of T. After relabeling we may assume that the edge e separates the leaves $1, 2, \ldots, i$ from the leaves $i+1, \ldots, n$. We construct two matrices M^e_{even} and M^e_{odd} each having 2^{i-1} rows and 2^{n-i-1} columns. The rows of M^e_{even} are indexed by subsets $I \subset \{1, \ldots, i\}$ with $|I|$ even and the columns are indexed by subsets $J \subset \{i+1, \ldots, n\}$ with $|J|$ even. The entry of M^e_{even} in row I and column J is the unknown $x_{I \cup J}$. The matrix M^e_{odd} is defined similarly. Our Gröbner basis for the toric variety X_T consists of the 2×2-minors of the matrices M^e_{even} and M^e_{odd} where e runs over all $2n-3$ edges of the tree T. In light of Theorem 3, it would be interesting to decide whether all the X_T lie on the same irreducible component of the Hilbert scheme, and, if yes, to explore possible connections between the generic point on that component to the quadratic equations derived by Keel and Tevelev [17] for the moduli space $\bar{M}_{0,n}$.

The toric variety X_T is known to evolutionary biologists as the *Jukes-Cantor model*. For some applications, it is more natural to study the *general Markov model*. This is a non-toric projective variety in tensor product space which generalizes secant varieties of Segre varieties [18]. The state of the art on the algebraic geometry of these models appears in the work of Elizabeth Allman and John Rhodes [1].

For our last theorem, we leave the field of phylogenetics and turn to mathematical developments inspired by other problems in biological sequence analysis. These problems include *gene prediction*, which seeks to identify genes inside genomes, and *alignment*, which aims to find the biological relationships between two genomes. See [22, §4] for an introduction aimed at mathematicians. Current algorithms for *ab initio* gene prediction and alignment are based on methods from statistical learning theory, and they involve *hidden Markov models* and more general *graphical models*.

From the perspective of algebraic statistics [21], a graphical model is a highly structured polynomial map from a low-dimensional space of parameters to a tensor product space, like the $\mathbb{P}^{2^{n-1}-1}$ we encountered in Theorem 3. It is from this algebraic representation of graphical models that the following theorem was derived:

Theorem 4. (Elizalde-Woods' Few Inference Functions) [11, 12]**)** *Consider a graphical model G with d parameters, where d is fixed, and let E be the number of edges of G. Then the number of inference functions of the model is at most $O(E^{d(d-1)})$.*

We need to explain what an inference function is and what this theorem means. A graphical model is given by a polynomial map $p : \mathbb{R}^d \to \mathbb{R}^N$ where d is fixed and each coordinate p_i is a polynomial of degree $O(E)$ in d unknown parameters. The polynomial p_i represents the probability of making the i-th observation $\#i$, out of a total of N possible observations. The number N is allowed to grow, and in biological applications it can be very large, for instance $N = 4^{1,000,000}$, the number of DNA sequences with one million base pairs.

The monomials in p_i correspond to the possible *explanations* of this observation, where the monomial of largest numerical value will be the most likely explanation. Let Exp be the set of all possible explanations for all the N observations. For a fixed generic choice of parameters $\theta \in \mathbb{R}^d$, we obtain a well-defined function

$$\phi_\theta \,:\, \{1, 2, \ldots, N\} \to \text{Exp}$$

which assigns to each observation its most likely explanation. Any such function, as θ ranges over (a suitable open subset of) $\mathbb{R}^d$ is called an *inference function* for the model f. The number $|\text{Exp}|^N$ of all conceivable functions is astronomical. The result by Elizalde and Woods says that only a tiny, tiny fraction of all these functions are actual inference functions. The polynomial growth rate in Theorem 4 makes it feasible, at least in principle, to pre-compute all such inference functions ahead of time, once per graphical model. This is important for *parametric inference*. Two recent examples of concrete bio-medical applications of parametric inference can be found in [3] and [7]. One way you can tell a biology paper from a mathematics paper is that the order of the authors' names has a meaning and is thus rarely alphabetic.

This concludes my discussion of four recent theorems that were inspired by biology. All four stem from my own limited field of expertise, and hence the selection has been very biased. A feature that Theorems 1, 2, 3 and 4 have in common is that they are meaningful as statements of pure mathematics. I must sincerely apologize to my colleagues in mathematical biology for having failed to give proper credit to their many many important research contributions. My only excuse is the hope that they will agree with my view that the answer to the question in the title is affirmative.

References

[1] E. Allman and J. Rhodes: *Phylogenetic ideals and varieties for the general Markov model*, `math.AG/0410604`.

[2] H-J Bandelt and A. Dress: *A canonical decomposition theory for metrics on a finite set*, Advances in Mathematics **92** (1992) 47–105.

[3] N. Beerenwinkel, C. Dewey and K. Woods: *Parametric inference of recombination in HIV genomes*, `q-bio.GN/0512019`.

[4] L. Billera, S. Holmes and K. Vogtman: *Geometry of the space of phylogenetic trees*, Advances in Applied Mathematics **27** (2001) 733-767.

[5] W. Buczynska and J. Wisniewski: *On phylogenetic trees - a geometer's view*, `math.AG/0601357`.

[6] J.E. Cohen: *Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better*, PLOS Biology **2** (2004) No.12.

[7] C. Dewey, P. Huggins, K. Woods, B. Sturmfels and L. Pachter: *Parametric alignment of Drosophila genomes*, PLOS Comput. Biology **2** (2006) No. 6.

[8] M. Déza and M. Laurent: *Geometry of Cuts and Metrics*, Springer, New York, 1997.

[9] A. Dress, K. Huber and V. Moulton: *Metric spaces in pure and applied mathematics*, Documenta Mathematica, Quadratic Forms LSU (2001) 121-139.
[10] A. Dress and W. Terhalle: *The tree of life and other affine buildings*, Documenta Mathematica, Extra Volume ICM III (1998) 565-574
[11] S. Elizalde: *Inference functions*, Chapter 9 in [21], pp. 215–225.
[12] S. Elizalde and K. Woods: *Bounds on the number of inference functions of a graphical model*, Formal Power Series and Algebraic Combinatorics (FPSAC 18), San Diego, June 2006.
[13] N. Eriksson, K. Ranestad, B. Sturmfels and S. Sullivant: *Phylogenetic algebraic geometry*, in Projective Varieties with Unexpected Properties, (editors C. Ciliberto, A. Geramita, B. Harbourne, R-M. Roig and K. Ranestad), De Gruyter, Berlin, 2005, pp. 237-255.
[14] M. Joswig: *Tight spans*, Introduction with link to the software `POLYMAKE` and an example of six bees, `www.mathematik.tu-darmstadt.de/∼joswig/tightspans/index.html`.
[15] D. H. Huson and D. Bryant: *Application of phylogenetic networks in evolutionary studies* `Molecular Biology and Evolution` **23** (2006) 254-267. (Software at `www.splitstree.org`)
[16] M. Kac, G-C. Rota and J. T. Schwartz: *Discrete Thoughts*, Birkhäuser, Boston,1986.
[17] S. Keel and J. Tevelev: *Equations for* $\bar{M}_{0,n}$, `math.AG/0507093`.
[18] JM Landsberg and L. Manivel: *On the ideals of secant varieties of Segre varieties*, Found Comput. Math. **4** (2004) 397-422
[19] D. Levy, R. Yoshida and L. Pachter: *Beyond pairwise distances: neighbor joining with phylogenetic diversity estimates*, Molecular Biology and Evolution **23** (2006) 491–498.
[20] L. Pachter and D. Speyer: *Reconstructing trees from subtree weights*, Applied Mathematics Letters **17** (2004) 615–621.
[21] L. Pachter and B. Sturmfels (eds.): *Algebraic Statistics for Computational Biology*, Cambridge University Press, 2005.
[22] L. Pachter and B. Sturmfels: *The mathematics of phylogenomics*, SIAM Review, to appear in 2007, `math.ST/0409132`.
[23] C. Semple and M. Steel: *Phylogenetics*, Oxford University Press, 2003.
[24] D. Speyer and B. Sturmfels: *The tropical Grassmannian*; Advances in Geometry **4** (2004), 389–411.
[25] B. Sturmfels and S. Sullivant: *Toric ideals of phylogenetic invariants*, Journal of Computational Biology **12** (2005) 204-228.

Department of Mathematics, Univ. of California, Berkeley CA 94720, USA

E-mail address: `bernd@math.berkeley.edu`