17

Extending Tree Models to Split Networks

David Bryant

17.1 Introduction

In this chapter we take statistical models designed for trees and adapt them for *split networks*, a more general class of mathematical structures. The models we propose provide natural swing-bridges between trees, filling in gaps in the probability simplex. There are many reasons why we might want to do this. Firstly, the split networks provide a graphical representation of phylogenetic uncertainty. Data that is close to tree-like produces a network that is close to a tree, while noisy or badly modeled data produce complex networks. Secondly, models that incorporate several trees open up possibilities for new tests to assess the relative support for different trees, in both likelihood and Bayesian frameworks. Thirdly, by searching through network space rather than tree space we may well be able to avoid some of the combinatorial headaches that make searching for trees so difficult.

17.2 Trees, splits and split networks

Splits are the foundation of phylogenetic combinatorics, and they will be the building blocks of our general statistical model. Recall (from Chapter 2) that a *split* $S = \{A, B\}$ of a finite set X is an unordered partition of X into two non-empty blocks. A *phylogenetic tree* for X is a pair $\mathcal{T} = (T, \phi)$ such that T is a tree with no vertices of degree two and ϕ is a bijection from X to the leaves of T.

Removing an edge e from an X-tree divides the tree into two connected components, thereby inducing a split of X that we say is the *split associated to* e. We use splits($\mathcal{T}$) to denote the sets associated to edges of T. The phylogenetic tree $\mathcal{T}$ can be reconstructed from the collection splits($\mathcal{T}$). The *Splits Equivalence Theorem* (Theorem 2.34) tells us that a collection $\mathcal{S}$ of splits is contained in splits($\mathcal{T}$) for some phylogenetic tree $\mathcal{T}$ if and only if the collection is *pairwise compatible*, that is, for all pairs of splits $\{A, B\}, \{A', B'\}$ at least

one of the intersections

$$A \cap A',\ A \cap B',\ B \cap A',\ B \cap B'$$

is empty.

If we think of phylogenetic trees as collections of compatible splits then it becomes easy to generalize trees: we simply consider collections of splits that are not necessarily pairwise compatible. This is the approach taken by Split Decomposition [Bandelt and Dress, 1992], Median Networks [Bandelt *et al.*, 1995], SpectroNet [Huber *et al.*, 2002], Neighbor-Net [Bryant and Moulton, 2004], Consensus Networks [Holland *et al.*, 2004] and Z-networks [Huson *et al.*, 2004], many of which are implemented in SplitsTree4 [Huson and Bryant, 2005]. The usefulness of these methods is due to a particularly elegant graphical representation for general collections of splits: the splits network.

To define splits networks, we first need to discuss splits graphs. These graphs have multiple characterizations. We will work with three of these here.

For a graph G let d_G denote the (unweighted) shortest path metric. A map ψ from a graph H to a graph G is an *isometric embedding* if $d_H(u,v) = d_G(\psi(u),\psi(v))$ for all $u, v \in V(H)$. A graph G is a *partial cube* if there exists an isometric embedding from G to a hypercube. [Wetzel, 1995] called these graphs *splits graphs*. This terminology has persisted in the phylogenetics community, despite the potential for confusion with the graph-theoretic term 'split graph' (a special class of perfect graphs). Refer to [Imrich and Klavžar, 2000] for a long list of characterizations for partial cubes.

[Wetzel, 1995] (see also [Dress and Huson, 2004]) characterized splits graphs in terms of isometric colorings. Let κ be an edge coloring of the graph. For each pair $u, v \in V(G)$ let $C_\kappa(u,v)$ denote the set of colors that appear on *every* shortest path from u to v. We say that κ is an *isometric coloring* if $d_G(u,v) = |C_\kappa(u,v)|$ for all pairs $u, v \in V(G)$. In other words, κ is isometric if the edges along any shortest path all have different colors, while any two shortest paths between the same pair of vertices have the same set of edge colors. A connected graph is a splits graph if and only if it has an isometric coloring [Wetzel, 1995].

A third characterization of splits graphs is due to [Winkler, 1984]. We define a relation Θ on pairs of edges $e_1 = \{u_1, v_1\}$ and $e_2 = \{u_2, v_2\}$ in a graph G by

$$e_1 \Theta e_2 \Leftrightarrow d_G(u_1,u_2) + d_G(v_1,v_2) \neq d_G(u_1,v_2) + d_G(v_1,u_2). \tag{17.1}$$

This relation is an equivalence relation if and only if G is a splits graph.

Two edges e_1 and e_2 in a splits graph have the same color in an isometric coloring if and only if the isometric embedding of the splits graph into the hypercube maps e_1 and e_2 to parallel edges, if and only if $e_1 \Theta e_2$. Thus, a splits graph has, essentially, a unique isometric coloring and a unique isometric

embedding into the hypercube. The partition of edges into color classes is completely determined by the graph.

Suppose now that we have a splits graph G and a map $\phi\colon X \to V(G)$. Using the isometric embedding, one can quickly prove that removing all edges in a particular color class partitions the graph into exactly two connected (and convex) components. This in turn induces a split of X, via the map ϕ. A *splits network* is a pair $\mathcal{N} = (G, \phi)$ such that

(i) G is a splits graph.
(ii) Each color class induces a distinct split of X.

The set of splits induced by the different color classes is denoted splits($\mathcal{N}$).

It is time for two examples. The split network on the left of Figure 17.1 corresponds to a collection of compatible splits - it is a tree. In this network, every edge is in a distinct color class. If we add the split $\{\{2,6\},\{1,3,4,5\}\}$ we obtain the split network on the right. There are four color classes in this graph that contain more than a single edge. These are the three horizontal pairs of parallel edges and the four edges marked in bold that induce the extra split.

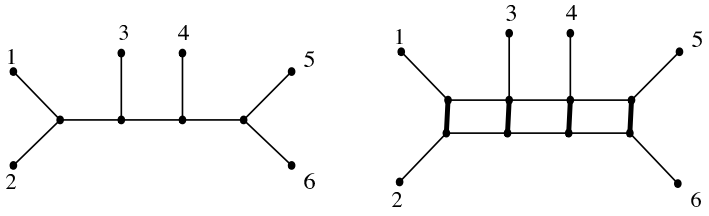

Fig. 17.1. Two splits networks. On the left, a split network for compatible splits (i.e. a tree). On the right, the same network with the split $\{\{2,6\},\{1,3,4,5\}\}$ included.

It is important to realize that the split network for a collection of splits may not be unique. Figure 17.2 reproduces an example in [Wetzel, 1995]. Both graphs are split networks for the set

$$\mathcal{S} = \Big\{\{\{1,2,3\},\{4,5,6,7\}\},\{\{2,3,4\},\{1,5,6,7\}\},\{\{1,2,7\},\{3,4,5,6\}\},\{\{1,2,6,7\},\{3,4,5\}\}\Big\}.$$

Each is minimal, in the sense that no subgraph of either graph is also a splits network for $\mathcal{S}$. In both graphs, the edges in the color class inducing the split $\{\{1,2,3\},\{4,5,6,7\}\}$ are in bold. In this example the two minimal graphs are isomorphic, but this is generally not the case.

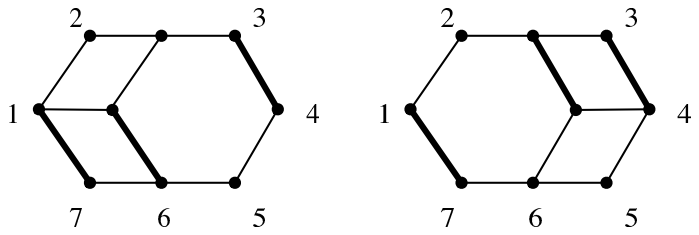

Fig. 17.2. Two different, and minimal, split networks for the same set of splits

17.3 Distance based models for trees and splits graphs

In molecular phylogenetics, the length of an edge in a tree is typically measured in terms of the average (or expected) number of mutations that occurred, per site, along that edge. The *evolutionary distance* between two sequences equals the sum of the lengths of the edges along the unique path the connects them in the unknown 'true' phylogeny. There is a host of methods for estimating the evolutionary distance starting from the sequences alone. These form the basis of distance based approaches to phylogenetics.

The oldest statistical methods for phylogenetics use models of how evolutionary distances estimated from pairwise comparisons of sequences differ from the true evolutionary distances (or *phyletic distances*) in the true, but unknown, phylogenetic tree [Cavalli-Sforza and Edwards, 1967, Farris, 1972, Bulmer, 1991]. It is assumed that the pairwise estimates are distributed, at least approximately, according to a multi-variate normal density centered on the true distances. The variance-covariance matrix for the density, here denoted by $\mathbf{V}$, can be estimated from the data [Bulmer, 1991, Susko, 2003], though early papers used a diagonal matrix, or the identity, for $\mathbf{V}$.

Once we have a variance-covariance matrix, and the observed distances, we can begin maximum likelihood estimation of the true distances δ_T, from which we can construct the maximum likelihood tree. Note that the term maximum likelihood here refers only to our approximate distance based model, not to the maximum likelihood estimation introduced by [Felsenstein, 1981]. Let n be the number of leaves. The maximum likelihood estimator is the tree metric $\widehat{\delta_T}$ that maximizes the likelihood function

$$L(\widehat{\delta_T}) = \Phi_{\binom{n}{2}}(d - \delta_T|\mathbf{V})$$

where Φ_m is the probability density function for the m dimensional multivariate normal:

$$\Phi_m(x|\mathbf{V}) = \frac{1}{(2\pi)^{\frac{m}{2}}\sqrt{\det(\mathbf{V})}} e^{-\frac{1}{2}x^T\mathbf{V}^{-1}x}.$$

Equivalently, we can minimize the least squares residue

$$\sum_{w<x}\sum_{y<z}\big(\widehat{\delta_T}(w,x)-d(w,x)\big)\mathbf{V}^{-1}_{(wx)(yz)}\big(\widehat{\delta_T}(y,z)-d(y,z)\big).$$

In either formulation, the optimization is carried out over all tree metrics in $\mathcal{T}_X$, the space of X-trees (Chapter 2).

We can describe tree metrics in terms of linear combinations of split metrics. The split metric for a split $\{A,B\}$ is the pseudo-metric on X given by

$$\delta_{\{A,B\}}(x,y)=\begin{cases}0 & \text{if } \{x,y\}\subseteq A \text{ or } \{x,y\}\subseteq B;\\ 1 & \text{otherwise.}\end{cases}$$

Let $w_{\{A,B\}}$ denote the length of the edge associated to a split $\{A,B\}\in \text{splits}(T)$. Then

$$\delta_T=\sum_{\{A,B\}\in\text{splits}(T)} w_{\{A,B\}}\delta_{\{A,B\}}. \tag{17.2}$$

This formulation can be used to estimate edge lengths on a fixed topology.

Equation (17.2) generalizes immediately to split networks. Suppose that the lengths of the edges in a split network $\mathcal{N}$ are given by the split weights $w_{\{A,B\}}$. Hence, all edges in the same color class have the same length. The distance between two labeled vertices x,y is the length of the shortest path between them, which in turn equals the sum of the weights of the splits separating x and y. We can therefore define a *network metric* $\mathcal{N}$ by

$$\delta_{\mathcal{N}}=\sum_{\{A,B\}\in\text{splits}(\mathcal{N})} w_{\{A,B\}}\delta_{\{A,B\}}.$$

The statistical model for distances from splits networks then works exactly as it did for phylogenetic trees. We assume that the observed distances d are distributed according to a multi-variate normal centered on the network metric $\delta_{\mathcal{N}}$. The covariance matrix can be estimated using the non-parametric method of [Susko, 2003]. The likelihood of a network metric $\widehat{\delta_{\mathcal{N}}}$ is, as before, given by $L(\widehat{\delta_{\mathcal{N}}})=\Phi_{\binom{n}{2}}(d-d_{\mathcal{N}}|\mathbf{V})$.

We immediately encounter the problem of identifiability. Phylogenetic trees, together with their edge lengths, are determined uniquely from their tree metrics. The same does not apply for network distances. The split metrics $\delta_{\{A,B\}}$ associated to splits of a network will not, in general, be linearly independent.

In practice, identifiability has not been too much of a problem. Split decomposition produces *weakly compatible* collections of splits. These have linearly independent split metrics and are uniquely determined from their network metrics [Bandelt and Dress, 1992]. Neighbor-Net produces networks based on

circular collections of splits which, as a subclass of weakly compatible splits, are also uniquely determined from their network metrics.

However the most important shortcoming of distance based methods, for either trees or networks, is that they lack the statistical efficiency of likelihood methods based on full stochastic models (see, e.g. [Felsenstein, 2003]). When we estimate distances from pairwise sequence comparisons we are effectively ignoring the joint probabilities of larger sets of sequences. What we gain in speed, we lose in accuracy.

17.4 A graphical model on a splits network?

The Markov model for trees outlined in Chapter 2 and Chapter 4 is just a special case in a general class of *graphical models*. Given the vast literature on graphical models, it seems that the logical generalization of the hidden tree model would be a graphical model defined on the splits network. This was the approach taken by [Strimmer and Moulton, 2000, Strimmer *et al.*, 2001].

Let $\mathcal{N}$ be a splits network. The first step is to choose a root and direct all edges away from the root (Figure 17.3). We now can apply a directed graphical model. The probability that a node is assigned a particular state depends on the states assigned to its parents: Strimmer and Moulton suggest several ways that this may be done.

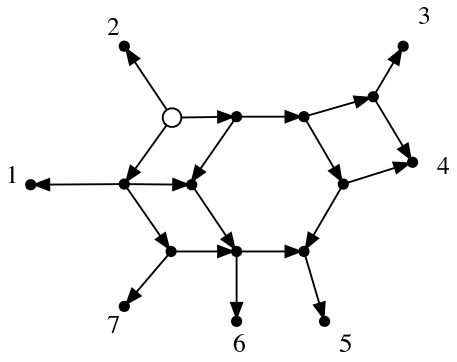

Fig. 17.3. Edge directions induced by placing the root at the white vertex.

There are significant problems with this general approach. Firstly, the probability of observing the data changes for different positions of the root, even when the mutation process is a time reversible model. It was claimed that this permitted estimation of the root, but there is no indication that the differences in distributions corresponded to any evolutionary phenomenon.

Secondly, different split networks for the *same* set of splits give different

pattern probabilities, even though the networks represent exactly the same information.

Thirdly, the internal nodes in split networks do not represent hypothetical ancestors. They are products of an embedding in a hypercube.

Strimmer *et al.* eventually concluded that split networks may not provide a suitable underlying graph for a stochastic network [Strimmer *et al.*, 2001]. It is true that graphical model technology can not be applied 'straight-off-the-shelf' to split networks. We need to be more sensitive to the particular properties of split networks. In the following section we will develop a model for split networks that avoids the problems encountered in this graphical model approach. The downside, however, is that we must first restrict ourselves to a special class of mutation models: group based models.

17.5 Group based mutation models

A mutation model on state space $\{1, 2, \ldots, r\}$ is said to be a *group based model* if there exists an Abelian group G with elements $g_1, \ldots, g_r$ and a function $\psi\colon G \to \mathbb{R}$ such that the instantaneous rate matrix Q satisfies

$$Q_{ij} = \psi(g_j - g_i)$$

for all i, j. The group operation on G is denoted using addition and we will use 0 for the identity element.

Let f be a homomorphism from G to the multiplicative group if complex numbers with modulus one. Thus $f(g + g') = f(g)f(g')$ for all $g, g' \in G$. The set of these homomorphisms forms a group $\widehat{G}$ that is isomorphic to G. We label the elements of $\widehat{G}$ so that the map $g \mapsto \widehat{g}$ taking $g \in G$ to $\widehat{g} \in \widehat{G}$ is an isomorphism. If $g = 0$ then $\widehat{g}$ is the function taking every element of G to 1. As usual, the conjugate of a complex number x is written $\overline{x}$.

Lemma 17.1 *Suppose that $g, h, h' \in G$, $a \in \mathbb{Z}$. Then we have the following identities:*

$$\begin{aligned}
\widehat{g}(-h) &= \overline{\widehat{g}(h)};\\
\widehat{g}(h + h') &= \widehat{g}(h)\widehat{g}(h');\\
\widehat{(h + h')}(g) &= \widehat{h}(g)\widehat{h'}(g);\\
\widehat{ag}(h) &= \widehat{g}(ah);
\end{aligned}$$

as well as the orthogonality property

$$\sum_{h \in G} \widehat{g}(h) = \begin{cases} |G| & \text{if } g = h;\\ 0 & \text{otherwise.}\end{cases}$$

Proof See, for example, [Körner, 1989]. □

Some indexing conventions will make our life easier. Since the elements of G are in one to one correspondence with $\{1, 2, \ldots, r\}$ we will index Q and $P(t)$ by group elements. So $Q_{g_i g_j}$ is equivalent to Q_{ij}.

We start with some basic observations about group based models.

Lemma 17.2 (i) *The eigenvalues of Q are given by*

$$\lambda_g = \sum_{h \in G} \overline{\widehat{g}(h)} \psi(h).$$

(ii) *The transition probabilities are given by*

$$P_{gg'}(t) = \frac{1}{r} \sum_{h \in G} \widehat{h}(g' - g) e^{\lambda_h t}.$$

(iii) *The uniform distribution is a stationary distribution.*

(iv) *If the process is ergodic and time reversible then $\psi(g) = \psi(-g)$ for all $g \in G$.*

Proof Define the $r \times r$ matrix K by $K_{ij} = \widehat{g}_i(g_j)$. Then

$$\begin{aligned}
(KQ)_{gg'} &= \sum_{h \in G} \widehat{g}(h) \psi(g' - h) \\
&= \sum_{h \in G} \widehat{g}(g' - h) \psi(h) \qquad \text{[replacing } h \text{ by } g' - h\text{]} \\
&= \widehat{g}(g') \sum_{h \in G} \overline{\widehat{g}(h)} \psi(h) \\
&= K_{gg'} \lambda_g.
\end{aligned}$$

Thus the rows of K are left-eigenvectors for Q. This proves (i). Let Λ be the diagonal matrix with $\Lambda_{gg} = \lambda_g$. Then $Q = K^{-1} \Lambda K$. By the orthogonality property in Lemma 17.1 we have $K^{-1} = \frac{1}{|G|} K^*$. Thus

$$\begin{aligned}
P_{gg'}(t) &= (e^{Qt})_{gg'} \\
&= \frac{1}{|G|} (K^* e^{\Lambda t} K)_{gg'} \\
&= \frac{1}{r} \sum_{h \in G} \overline{\widehat{h}(g)} e^{\lambda_h t} \widehat{h}(g') \\
&= \frac{1}{r} \sum_{h \in G} \widehat{h}(g' - g) e^{\lambda_h t}
\end{aligned}$$

proving (ii). For (iii), observe that the first row of K gives a left-eigenvector that is all ones. Finally, if the process is ergodic then the uniform distribution

is the unique stationary distribution. This, together with the assumption that the process is time reversible, implies that both Q and $P(t)$ are symmetric and that $\psi(g) = \psi(-g)$ for all g. □

As an example, consider the case when $r = 4$. There are two (up to isomorphism) Abelian groups on four elements, $\mathbb{Z}_4$ and $\mathbb{Z}_2 \times \mathbb{Z}_2$. If $G = \mathbb{Z}_4$ then the condition that $\psi(g) = \psi(-g)$ implies that Q must have the form

$$Q = \begin{pmatrix} -2a-b & a & b & a \\ a & -2a-b & a & b \\ b & a & -2a-b & a \\ a & b & a & -2a-b \end{pmatrix}$$

which is Kimura's two parameter (K2P) model (Chapter 4). If $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ then we always have $g = -g$ so there are three parameters available for Q:

$$Q = \begin{pmatrix} -a-b-c & a & b & c \\ a & -a-b-c & c & b \\ b & c & -a-b-c & a \\ c & b & a & -a-b-c \end{pmatrix}.$$

In this case we obtain Kimura's three parameter model (K3P) [Kimura, 1981].

17.6 Group based models for trees and splits

Suppose that we have an ergodic, time reversible, group based mutation model with state set $\Sigma = \{1, 2, \ldots, r\}$ and Abelian group G, $|G| = r$, where $Q_{ij} = \psi(g_j - g_i)$ for all i, j. Let $P(t) = e^{Qt}$ denote the corresponding transition probabilities. Let $\mathcal{T} = (T, \phi)$ be a phylogenetic tree with n leaves. We use $t_e = t_{kl}$ to denote the length of an edge $e = kl \in E(T)$. In terms of the tree model of Chapter 2, $\theta^{kl} = P(t_{kl})$ for all $kl \in E(T)$.

We define

$$\rho_t(g) = \frac{1}{r} \sum_{h \in G} \widehat{h}(g) e^{\lambda_h t}$$

so that by Lemma 17.2, $P_{gg'}(t) = \rho_t(g' - g)$ for all $g, g' \in G$ and $t \geq 0$.

Lemma 17.3 *Let σ be a map from $V(T)$ to Σ. For each edge $e = kl$ define $x_e = g_{\sigma_l} - g_{\sigma_k}$. Then*

$$p_\sigma = \frac{1}{r} \prod_{e \in E(T)} \rho_{t_e}(x_e).$$

Proof By Lemma 17.2 the mutation model has a uniform stationary distribution. We can therefore apply (1.53), giving

$$\begin{aligned} p_\sigma &= \frac{1}{|\Sigma|} \prod_{kl \in E(T)} \theta^{kl}_{\sigma_k \sigma_l} \\ &= \frac{1}{r} \prod_{kl \in E(T)} \rho_{t_{kl}}(g_{\sigma_l} - g_{\sigma_k}) \\ &= \frac{1}{r} \prod_{e \in E(T)} \rho_{t_e}(x_e). \end{aligned}$$

□

Let χ be a map from the leaves of T to Σ. We say that $\sigma : V(T) \to \Sigma$ extends χ if $\sigma_i = \chi_i$ for all leaves i. Under the *hidden tree model* the probability of observing χ is defined

$$p_\chi = \sum_{\sigma : \sigma \text{ extends } \chi} p_\sigma.$$

Suppose that $E(T) = \{e_1, e_2, \ldots, e_q\}$, let $\{A_k, B_k\}$ be the split associated to edge k and let $\mathbf{A}$ be the $(n-1) \times q$ matrix defined by

$$\mathbf{A}_{ik} = \begin{cases} 1 & i \text{ and } n \text{ are on opposite sides of } \{A_k, B_k\} \\ 0 & \text{otherwise.} \end{cases} \tag{17.3}$$

The next observation is crucial, since it allow use to re-express the likelihood in a form that extends immediately to arbitrary collections of splits.

Theorem 17.4 *Define the vector $y = y[\chi] \in G^{n-1}$ by $y_i = g_{\chi_n} - g_{\chi_i}$. Let σ be a map from $V(T)$ to Σ. For each edge $e = kl$ define $x_e = g_{\sigma_l} - g_{\sigma_k}$. Then σ extends χ if and only if $\mathbf{A}x = y$. Furthermore, the probability of observing χ is given by*

$$p_\chi = \sum_{x \in G^q : \mathbf{A}x = y} \prod_{e \in E(T)} \rho_{t_e}(x_e). \tag{17.4}$$

Proof We prove that $\mathbf{A}x = y$ if and only if σ extends χ. The second claim then follows from Lemma 17.3.

For each leaf i, let E_i be the edges on the path from leaf n to leaf i. We will assume that T is rooted at leaf n, so all edges in E_i are directed away from n.

Then

$$\begin{aligned}(\mathbf{A}x)_i &= \sum_{kl \in E_i} x_{kl} \\ &= \sum_{kl \in E_i} (g_{\sigma_l} - g_{\sigma_k}) \\ &= g_{\sigma_n} - g_{\sigma_i}.\end{aligned}$$

Thus $\mathbf{A}x = y$ if and only if $g_{\sigma_i} = g_{\chi_i}$ for all leaves i, if and only if $\sigma_i = \chi_i$ for all leaves i, if and only if σ extends χ. □

The importance of Theorem 17.4 so far as we are concerned is that p_χ is not expressed in terms of the tree structure: it is defined in terms of splits. We can therefore generalize the definition of pattern probabilities to any collection of splits.

Let $\mathcal{N}$ be a weighted split network with splits $\{A_1, B_1\}, \{A_2, B_2\}, \ldots, \{A_q, B_q\}$ and let t_k be the length assigned to split $\{A_k, B_k\}$. Let A be the matrix defined by (17.6). The *probability of a phylogenetic character* χ *given* $\mathcal{N}$ is then defined by

$$p_\chi = \sum_{x \in G^q : \mathbf{A}x = y} \prod_{k=1}^{q} \rho_{t_k}(x_k). \tag{17.5}$$

The uncanny similarity between (17.4) and (17.5) now gives

Theorem 17.5 *Let $\mathcal{N}$ be a weighted split network. If the splits of $\mathcal{N}$ are compatible then the character probabilities correspond to exactly those given by the tree based model.*

We can rephrase this model in terms of graphical models on the splits network. We say that a map $\sigma : V(\mathcal{N}) \to \Sigma$ is *concordant* if $\sigma_l - \sigma_k = \sigma_j - \sigma_i$ for all pairs of edges $ij, kl \in E(\mathcal{N})$ in the same color class. The probability of a map σ is just the product of $P_{\sigma_k \sigma_l}(t_{kl})$ over all edges $kl \in E(T)$, where t_{kl} is the length of the edge. We then have that p_χ equals the probability that a map σ extends χ, conditional on σ being concordant.

17.7 A Fourier calculus for split networks

[Székely *et al.*, 1993] describe a *Fourier calculus on evolutionary trees* that generalizes the Hadamard transform of [Hendy and Penny, 1989, Steel *et al.*, 1992]. Using their approach, we can take the observed character frequencies, apply a transformation, and obtain a vector of values from which we can read off the support for different splits. They show that if the observed character frequencies correspond exactly to the character probabilities determined by some

phylogenetic tree then the split supports will correspond exactly to the splits and branch lengths in the phylogenetic tree. Conversely, the inverse transformation gives a single formula for the character probabilities in any tree.

This theory generalizes seamlessly from trees to split networks—in fact so seamlessly that the proofs of [Székely *et al.*, 1993] require almost no modifications to establish the general case. Their approach was prove that their transform worked when applied to character probabilities from a tree. The correctness of the inverse formula then followed by applying a Fourier transformation. In this section, we will prove the same results but working in the opposite direction. We show that, starting with weights on the splits, a single invertible formula gives the character probabilities. Our motivation is that, at some point in the future, we will need to generalize these results beyond Abelian group based models, and the elegant Fourier inversion formula may not exist in this context.

For $x, y \in G^m$ we define

$$\widehat{y}(x) = \prod_{i=1}^{m} \widehat{y_i}(x_i).$$

The set $\{\widehat{y} : y \in G^m\}$ forms a group under multiplication that is isomorphic to G^m.

Lemma 17.6 *Suppose that $z \in G^q$ and $y \in G^{n-1}$. Let $\mathbf{A}$ be an $(n-1) \times q$ integer matrix. Either*

$$\sum_{x \in G^q : \mathbf{A}x = y} \widehat{z}(x) = 0$$

or there is $u \in G^{n-1}$ such that $z = \mathbf{A}^T u$ and so

$$\sum_{x \in G^q : \mathbf{A}x = y} \widehat{z}(x) = r^{q-(n-1)} \widehat{z}(u)$$

Proof Suppose that $\sum_{x \in G^q : \mathbf{A}x = y} \widehat{z}(x) \neq 0$. For any v such that $\mathbf{A}v = 0$ we have

$$\sum_{x \in G^q : \mathbf{A}x = y} \widehat{z}(x) = \sum_{x \in G^q : \mathbf{A}x = y} \widehat{z}(x+v) = \widehat{z}(v) \sum_{x \in G^q : \mathbf{A}x = y} \widehat{z}(x)$$

so $\widehat{z}(v) = 1$.

For every $x, y \in G^{n-1}$ we have

$$\mathbf{A}x = \mathbf{A}y \Leftrightarrow \mathbf{A}(x-y) = 0 \Leftrightarrow \widehat{z}(x-y) = 1 \Leftrightarrow \widehat{z}(x) = \widehat{z}(y).$$

Define $H = \{\mathbf{A}x : x \in G^q\}$, so that H forms a normal subgroup of G^{n-1}. Define the map $f : H \to \mathbb{C}$ by setting $f(\mathbf{A}x) = \widehat{z}(x)$ for all $x \in G^q$. This is a

homomorphism from H to the unit circle, since $f(\mathbf{A}x+\mathbf{A}y) = f(\mathbf{A}(x+y)) = \widehat{z}(x+y) = \widehat{z}(x)\widehat{z}(y) = f(\mathbf{A}x)f(\mathbf{A}y)$. By Lemma 104.3 of [Körner, 1989] we can extend f to the rest of G. Thus there is u such that $f = \widehat{u}$ and, for all $x \in G^q$, $\widehat{z}(x) = \widehat{u}(\mathbf{A}x)$. The result now follows by expanding $\widehat{u}(\mathbf{A}x)$. □

We now arrive at our main theorem. It provides the formula linking weights on splits to pattern probabilities, for trees *and* split networks.

Theorem 17.7 *Let $\mathcal{S}$ be a collection of splits, $|\mathcal{S}| = q$ and let $\mathbf{A}$ be the $(n-1)\times q$ matrix defined by*

$$\mathbf{A}_{ik} = \begin{cases} 1 & i \text{ and } n \text{ are on opposite sides of } \{A_k, B_k\} \\ 0 & \text{otherwise.} \end{cases} \tag{17.6}$$

Let b be the real valued vector with entries indexed by G^{n-1} so that for all $z \in G^{n-1}$,

$$b_z = \begin{cases} \psi(h)t_k & \text{if there is } h \in G \text{ and } k \text{ such that } z_i = h\mathbf{A}_{ik} \text{ for all } i \\ -\sum_{v\in G^{n-1}-\{0\}} b_v & \text{if } z = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathbf{H}$ be the matrix with rows and columns indexed by G^{n-1} and $\mathbf{H}_{gg'} = \widehat{g}(g)$.

Given any map χ from the leaves to the set of states $\{1, 2, \ldots, r\}$ define $y \in G^{n-1}$ by $y_i = g_{\chi_n} - g_{\chi_i}$ for all leaves $i = 1, \ldots, n-1$. Then the probability of χ is given by

$$p_\chi = \left[\mathbf{H}^{-1}\exp[\mathbf{H}b]\right]_y. \tag{17.7}$$

Proof From (17.5) we have

$$\begin{aligned} p_\chi &= \sum_{\substack{x:\mathbf{A}x=y \\ x\in G^q}} \prod_{k=1}^{q} \rho_{t_k}(x_k) \\ &= \sum_{\substack{x:\mathbf{A}x=y \\ x\in G^q}} \prod_{k=1}^{q} \frac{1}{r}\sum_{h\in G} \widehat{h}(x_k)e^{\lambda_h t_k} \\ &= \frac{1}{r^q}\sum_{\substack{x:\mathbf{A}x=y \\ x\in G^q}} \sum_{z\in G^q} \prod_{k=1}^{q} \widehat{z_k}(x_k)e^{\lambda_{z_k} t_k} \\ &= \frac{1}{r^q}\sum_{z\in G^q}\left(\sum_{\substack{x:\mathbf{A}x=y \\ x\in G^q}} \widehat{z}(x)\right)\exp\left[\sum_{k=1}^{q}\lambda_{z_k}t_k\right]. \end{aligned}$$

So far we have just applied the definitions, reversed a summation and product, and regrouped. From Lemma 17.6 we have that $\sum_{x \in G^q : \mathbf{A}x = y} \widehat{z}(x)$ equals zero unless $z = -\mathbf{A}^T u$ for some $u \in G^{n-1}$. We therefore ignore all z for which this does not hold. Substituting in and using $\widehat{-u} = \overline{\widehat{u}}$ we obtain

$$\begin{aligned} p_\chi &= \frac{1}{r^{n-1}} \sum_{u \in G^{n-1}} \overline{\widehat{u}}(y) e^{\beta_u} \\ &= \left[\mathbf{H}^{-1} \exp[\beta] \right]_y \end{aligned}$$

where

$$\begin{aligned} \beta_u &= \sum_{k=1}^{q} \lambda_{(\mathbf{A}^T \overline{u})_k} t_k \\ &= \sum_{k=1}^{q} \sum_{h \in G} \widehat{\mathbf{A}^T u_k}(h) \psi(h) t_k \\ &= \sum_{k=1}^{q} \sum_{h \in G} \widehat{u}(\eta_{kh}) \psi(h) t_k \\ &= \sum_{v \in G^{n-1}} \widehat{u}(v) b_v \\ &= \mathbf{H} b. \end{aligned}$$

□

We have proven, more or less, Theorem 6 of [Székely *et al.*, 1993] without any reference to trees. In the special case that $r = 2$, (17.7) becomes the classical *Hadamard transform* of [Hendy and Penny, 1989, Steel *et al.*, 1992]. This is comforting: [Felsenstein, 2003] describes the Hadamard type approach as "one of the nicest applications of mathematics to phylogenies so far."

Note that the formula $\mathbf{H}^{-1} \exp[\mathbf{H}b]$ is invertible. This means that every split network gives a different character distribution. We cannot recover split networks from their distance metrics $d_{\mathcal{N}}$ but we can recover them from their character probabilities. A maximum likelihood estimator based on (17.7) will be statistically consistent.

One key problem remains. The constraint that we only use group based mutation models is too much of a restriction. For nucleotide data, and especially for protein data, a uniform stationary distribution is unrealistic. It is reasonable to believe that some reasonable generalization of these results exists for more general mutation models: after all there is no such restriction on distance based methods. What exact form these generalizations will take is, at the moment, anybody's guess.

Bibliography

[Abril *et al.*, 2005] JF Abril, R Castelo, and R Guigó. Comparison of splice sites in mammals and chicken. *Genome Research*, 15:111–119, 2005.

[Adkins *et al.*, 2001] Ronald M. Adkins, Eric L. Gelke, Diane Rowe, and Rodney L. Honeycutt. Molecular Phylogeny and Divergence Time Estimates for Major Rodent Groups: Evidence from Multiple Genes. *Mol Biol Evol*, 18(5):777–791, 2001.

[Agresti, 1990] Alan Agresti. *Categorical data analysis*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1990. A Wiley-Interscience Publication.

[Aji and McEliece, 2000] Srinivas M Aji and Robert J McEliece. The generalized distributive law. *IEEE Trans. Inform. Theory*, 46(2):325–343, 2000.

[Alefeld and Herzberger, 1983] G Alefeld and J Herzberger. *An introduction to interval computations*. Academic press, New York, 1983.

[Allman and Rhodes, 2003] Elizabeth S Allman and John A Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Math. Biosci.*, 186(2):113–144, 2003.

[Allman and Rhodes, 2004] Elizabeth S Allman and John A Rhodes. Quartets and parameter recovery for the general Markov model of sequence mutation. *AMRX Appl. Math. Res. Express*, (4):107–131, 2004.

[Altschul *et al.*, 1990] SF Altschul, W Gish, W Miller, EW Myers, and DJ Lipman. Basic local alignment search tool. *J Molecular Biology*, 215:403–410, 1990.

[Apostol, 1976] Tom M Apostol. *Introduction to analytic number theory*. Springer-Verlag, New York, 1976. Undergraduate Texts in Mathematics.

[Ardila, 2004] F Ardila. A tropical morphism related to the hyperplane arrangement of the complete bipartite graph. *Discrete and Computational Geometry, to appear*, 2004.

[Aris-Brosou, 2003] S Aris-Brosou. How bayes tests of molecular phylogenies compare with frequentist approaches. *Bioinformatics*, 19:618–624, 2003.

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

[Bandelt *et al.*, 1995] H.J. Bandelt, P. Forster, B.C. Sykes, and M.B. Richards. Mitochondrial portraits of human population using median networks. *Genetics*, 141:743–753, 1995.

[Beerenwinkel *et al.*, 2004] N Beerenwinkel, J Rahnenführer, M Däumer, D Hoffmann, R Kaiser, J Selbig, and T Lengauer. Learning multiple evolutionary pathways from cross-sectional data. In *Proc. 8th Ann. Int. Conf. on Res. in Comput. Biol. (RECOMB '04), 27–31 March 2004, San Diego, CA*, pages 36–44, 2004. to appear in *J. Comp. Biol.*

[Beerenwinkel *et al.*, 2005a] N Beerenwinkel, M Däumer, T Sing, J Rahnenführer, T Lengauer, J Selbig, D Hoffmann, and R Kaiser. Estimating HIV evolutionary pathways and the genetic barrier to drug resistance. *J. Infect. Dis.*, 2005. to appear.

[Beerenwinkel *et al.*, 2005b] N Beerenwinkel, J Rahnenführer, R Kaiser, D Hoffmann, J Selbig, and T Lengauer. Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics*, 2005. to appear.

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

[Berz, 1991] M Berz. Forward algorithms for high orders and many variables with application to beam physics. In A Griewank and G Corliss, editors, *Automatic differentiation of algorithms: theory, implementation and applications*, pages 147–

156. SIAM, Philadelphia, PA, 1991.

[Boffelli *et al.*, 2003] Dario Boffelli, Jon McAuliffe, Dmitriy Ovcharenko, Keith D Lewis, Ivan Ovcharenko, Lior Pachter, and Edward M Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–4, 2003.

[Boffelli *et al.*, 2004] D. Boffelli, M.A. Nobrega, and E.M. Rubin. Comparative genomics at the vertebrate extremes. *Nature Reviews Genetics*, 5:456–465, 2004.

[Bosma *et al.*, 1997] Wieb Bosma, John Cannon, and Catherine Playoust. The MAGMA algebra system I: the user language. *J. Symb. Comput.*, 24(3-4):235–265, 1997.

[Bourque *et al.*, 2004] Guillaume Bourque, Pavel A Pevzner, and Glenn Tesler. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res*, 14(4):507–16, 2004.

[Bray and Pachter, 2004] Nicolas Bray and Lior Pachter. Mavid: constrained ancestral alignment of multiple sequences. *Genome Res*, 14(4):693–9, 2004.

[Brown *et al.*, 1982] WM Brown, EM Prager, A Wang, and AC Wilson. Mitochondrial DNA sequences of primates, tempo and mode of evolution. *Journal of Molecular Evolution*, 18:225–239, 1982.

[Brown, 2002] TA Brown. *Genomes 2*. John Wiley & Son, Inc., 2002.

[Bryant and Moulton, 2004] D. Bryant and V. Moulton. Neighbornet: An agglomerative algorithm for the construction of planar phylogenetic networks. *Molecular Biology And Evolution*, 21:255–265, 2004.

[Buchberger, 1965] B Buchberger. *An algorithm for finding a basis for the residue class ring of a zero-dimensional polynomial ideal (in German)*. PhD thesis, Univ. Innsbruck, Dept. of Math., Innsbruck, Austria, 1965.

[Bulmer, 1991] D. Bulmer. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Molecular Biology and Evolution*, 8(6):868–883, 1991.

[Catanese *et al.*, 2004] Fabrizio Catanese, Serkan Hoşten, Amit Khetan, and Bernd Sturmfels. The maximum likelihood degree. `math.AG/0406533`, 2004.

[Cavalli-Sforza and Edwards, 1967] L. Cavalli-Sforza and A. Edwards. Phylogenetic analysis models and estimation procedures. *Evolution*, 32:550–570, 1967.

[Cavender and Felsenstein, 1987] J. Cavender and J. Felsenstein. Invariants of phylogenies in a simple case with discrete states. *Journal of Classification*, 4:57–71, 1987.

[Chargaff, 1950] E Chargaff. Chemical specificity of nucleic acids and mechanism for the enzymatic degradation. *Experientia*, 6:201–209, 1950.

[Cohen, 2004] Joel E Cohen. Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biol*, 2(12):e439, 2004.

[Cox *et al.*, 1997] David Cox, John Little, and Donal O'Shea. *Ideals, varieties, and algorithms*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, second edition, 1997. An introduction to computational algebraic geometry and commutative algebra.

[Craciun and Feinberg, 2004] G Craciun and M Feinberg. Multiple equilibria in complex chemical reaction networks: I. the injectivity property. *SIAM Journal of Applied Mathematics*, 2004.

[Cuyt *et al.*, 2001] A Cuyt, B Verdonk, S Becuwe, and P Kuterna. A remarkable example of catastrophic cancellation unraveled. *Computing*, 66:309–320, 2001.

[Darwin, 1859] C Darwin. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray, London, 1859.

[Demmel, 1997] James W. Demmel. *Applied numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[Desper *et al.*, 1999] R. Desper, F. Jiang, O.-P. Kallioniemi, H. Moch, C.H. Papadimitriou, and A.A. Schäffer. Inferring tree models for oncogenesis from comparative genome hybridization data. *J. Comp. Biol.*, 6(1):37–51, 1999.

[Deza and Laurent, 1997] Michel Marie Deza and Monique Laurent. *Geometry of cuts and metrics*, volume 15 of *Algorithms and Combinatorics*. Springer-Verlag, Berlin, 1997.

[Douzery E. J. P. and D., 2003] Stanhope M. J. Douzery E. J. P., Delsuc F. and Huchon D. Local molecular clocks in three nuclear genes: divergence ages of rodents and other mammals, and incompatibility between fossil calibrations. *Molecular Biology and Evolution*, 57:201–213, 2003.

[Dress and Huson, 2004] A. Dress and Daniel Huson. Constructing splits graphs. *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, 2004.

[Dress and Terhalle, 1998] Andreas Dress and Werner Terhalle. The tree of life and other affine buildings. In *Proceedings of the International Congress of Mathematicians, Vol. III (Berlin, 1998)*, number Extra Vol. III, pages 565–574 (electronic), 1998.

[Durbin *et al.*, 1998] R Durbin, S Eddy, A Korgh, and G Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[Eichler and Sankoff, 2003] EE Eichler and D. Sankoff. Structural dynamics of eukaryotic chromosome evolution. *Science*, 301:793–797, 2003.

[ENCODE, 2004] The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–40, 2004.

[Farris, 1972] J.S. Farris. Estimating phylogenetic trees from distance matrices. *American Naturalist*, 106:645–668, 1972.

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

[Felsenstein, 2003] J Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Inc., 2003.

[Felsenstein, 2004] J Felsenstein. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author, Department of Genome Sciences, University of Washington, Seattle, 2004.

[Fleischmann *et al.*, 1995] RD Fleischmann, MD Adams, O White, RA Clayton, EF Kirkness, AR Kerlavage, CJ Bult, JF Tomb, BA Dougherty, and JM Merrick et al. Whole-genome random sequencing and assembly of *haemophilus influenza* Rd. *Science*, 269(5223):496–512, 1995.

[Floyd, 1962] RW Floyd. Algorithm 97: shortest path. *Communications of ACM*, 5(6):345, 1962.

[Forney, 1973] G.D. Forney. The viterbi algorithm. *Proc. of the IEEE*, 61(3):268–278, 1973.

[Galtier and Gouy, 1998] N. Galtier and M. Gouy. Inferring pattern and process: maximum likelihood implementation of a non-homogeneous model of dna sequence evolution for phylogenetic analysis. *Molecular Biology and Evolution*, 154(4):871–879, 1998.

[Garcia *et al.*, 2004] L. D. Garcia, M. Stillman, and B. Sturmfels. Algebraic geometry of Bayesian networks. *J. Symbolic Comput.*, 2004. Special Issue Méthodes Effectives en Géometrie Algébrique (MEGA).

[Garcia, 2004] L. D. Garcia. Algebraic statistics in model selection. In M. Chickering and J. Halpern, editors, *Proceedings of the* 20^{th} *Conference on Uncertainty in Artificial Intelligence*, pages 177–184. AUAI Press, Arlington, VA, 2004.

[Gawrilow and Joswig, 2000] Ewgenij Gawrilow and Michael Joswig. polymake: a

framework for analyzing convex polytopes. In Gil Kalai and Günter M. Ziegler, editors, *Polytopes — Combinatorics and Computation*, pages 43–74. Birkhäuser, 2000.

[Gawrilow and Joswig, 2001] Ewgenij Gawrilow and Michael Joswig. polymake: an approach to modular software design in computational geometry. In *Proceedings of the 17th Annual Symposium on Computational Geometry*, pages 222–231. ACM, 2001. June 3-5, 2001, Medford, MA.

[Geiger *et al.*, 2005] D. Geiger, C. Meek, and B. Sturmfels. On the toric algebra of graphical models. *Ann. Statist.*, to appear, 2005.

[Gentleman *et al.*, 2004] Robert C Gentleman, Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch Cheng Li, Martin Maechler, Anthony J. Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y. H. Yang, and Jianhua Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[Grayson and Stillman, 2002] Daniel R. Grayson and Michael E. Stillman. Macaulay 2, a software system for research in algebraic geometry. Available at http://www.math.uiuc.edu/Macaulay2/, 2002.

[Greuel and Pfister, 2002] GM Greuel and G Pfister. *A Singular Introduction to Commutative Algebra*. Springer-Verlag, Berlin and Heidelberg, 2002.

[Grünbaum, 2003] Branko Grünbaum. *Convex polytopes*, volume 221 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 2003. Prepared and with a preface by Volker Kaibel, Victor Klee and Günter M. Ziegler.

[Gusfield *et al.*, 1994] D Gusfield, K Balasubramanian, and D Naor. Parametric optimization of sequence alignment. *Algorithmica*, 12:312–326, 1994.

[Gusfield, 1997] D Gusfield. *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, 1997.

[Hammer *et al.*, 1995] R Hammer, M Hocks, U Kulisch, and D Ratz. *C++ toolbox for verified computing: basic numerical problems*. Springer-Verlag, Berlin, 1995.

[Hansen and Sengupta, 1981] E Hansen and S Sengupta. Bounding solutions of systems of equations using interval analysis. *BIT*, 21:203–211, 1981.

[Hansen, 1980] E Hansen. Global optimization using interval analysis - the multidimensional case. *Numerische Mathematik*, 34:247–270, 1980.

[Hansen, 1992] E Hansen. *Global optimization using interval analysis*. Marcel Dekker, New York, 1992.

[Hendy and Penny, 1989] M. Hendy and D. Penny. A framework for the quantitative study of evolutionary trees. *Systematic Zoology*, 38(4), 1989.

[Holland *et al.*, 2004] B. Holland, K.T. Huber, V. Moulton, and P. Lockhart. Using consensus networks to visualize contradictory evidence for species phylogeny. *Molecular Biology and Evolution*, 21(7):1459–1461, 2004.

[Huber *et al.*, 2002] K.T. Huber, M. Langton, V. Penny, D.and Moulton, and M. Hendy. Spectronet: A package for computing spectra and median networks. *Applied Bioinformatics*, 1(3):2041–2059, 2002.

[Huelsenbeck *et al.*, 2000] J. P. Huelsenbeck, B. Larget, and D. L. Swofford. A compound poisson process for relaxing the molecular clock. *Genetics*, 154(4):1879–1892, 2000.

[Human Genome Sequencing Consortium, 2004] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[Huson and Bryant, 2005] Daniel Huson and D. Bryant. Estimating phylogenetic trees and networks using splitstree4. 2005.

[Huson *et al.*, 2004] Daniel Huson, T. Dezulian, T. Kloepper, and M. Steel. Phylogenetic super-networks from partial trees. *IEEE Transactions on Computational Biology and Bioinformatics*, 1(4):151–158, 2004.

[Huson, 1998] D. Huson. Splitstree - a program for analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1):68–73, 1998.

[IEEE Task P754, 1985] IEEE, New York. *ANSI/IEEE 754-1985, Standard for Binary Floating-Point Arithmetic*, 1985. A preliminary draft was published in the January 1980 issue of IEEE Computer, together with several companion articles. Available from the IEEE Service Center, Piscataway, NJ, USA.

[Ihaka and Gentleman, 1996] Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[Imrich and Klavžar, 2000] Wilfried Imrich and Sandi Klavžar. *Product graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000. Structure and recognition, With a foreword by Peter Winkler.

[Jukes and Cantor, 1969] TH Jukes and C Cantor. Evolution of protein molecules. In HN Munro, editor, *Mammalian Protein Metabolism*, pages 21–32. New York Academic Press, 1969.

[Karlin and Altschul, 1990] S Karlin and SF Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences, USA*, 87:2264–2268, 1990.

[Kellis *et al.*, 2004] M. Kellis, B. Birren, and E. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisae. *Nature*, 8:617–624, 2004.

[Kent, 2002] James Kent. Blat- the blast like alignment tool. *Genome Biology*, 12(4):656–664, 2002.

[Kimura, 1981] M. Kimura. Estimation of evolutionary sequences between homologous nucleotide sequences. *Proc. Nat. Acad. Sci. U.S.A.*, 78:454–458, 1981.

[Korf *et al.*, 2003] I Korf, M Yandell, and J Bedell. *BLAST*. O'Reilly & Associates, Sebastopol, CA, 2003.

[Körner, 1989] T. W. Körner. *Fourier analysis*. Cambridge University Press, Cambridge, second edition, 1989.

[Kuhn, 1955] HW Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[Kulisch *et al.*, 2001] U Kulisch, R Lohner, and A Facius, editors. *Perspectives on encolsure methods*. Springer-Verlag, New York, 2001.

[Lake, 1987] J. A. Lake. A rate-independent technique for analysis of nucleaic acid sequences: evolutionary parsimony. *Molecular Biology and Evolution*, 4:167–191, 1987.

[Lander and Waterman, 1988] ES Lander and MS Waterman. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2:231–239, 1988.

[Landsberg and Manivel, 2004] J.M. Landsberg and L. Manivel. On the ideals of secant varieties of segre varieties. *Foundations of Computational Mathematics*, 4(4):397–422, 2004.

[Laubenbacher, 2003] R Laubenbacher. A computer algebra approach to biological systems. In *Proc. 2003 Intl. Symposium on Symbolic and Algebraic Computation*, 2003.

[Lauritzen, 1996] Steffen L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.

[Lenstra, 1983] H. W. Lenstra, Jr. Integer programming with a fixed number of vari-

ables. *Math. Oper. Res.*, 8(4):538–548, 1983.

[Levy *et al.*, 2004] D. Levy, R. Yoshida, and L. Pachter. Neighbor joining with subtree weights. *preprint*, 2004.

[Litvinov, 2005] G Litvinov. The maslov dequantization, idempotent and tropical mathematics: a very brief introduction. *arXiv.org:math/0501038*, 2005.

[Loh and Walster, 2002] E Loh and GW Walster. Rump's example revisited. *Reliable Computing*, 8:245–248, 2002.

[McMullen, 1971] P. McMullen. The maximum numbers of faces of a convex polytope. *J. Combinatorial Theory, Ser. B*, 10:179–184, 1971.

[Mindell and Honeycutt, 1990] D. P. Mindell and R. L. Honeycutt. Ribosomal rna in vertebrates: evolution and phylogenetic applications. *Ann. Rev. Ecol. Syst.*, 21:541–566, 1990.

[Mond *et al.*, 2003] DMQ Mond, JQ Smith, and D Van Straten. Stochastic factorisations, sandwiched simplices and the topology of the space of explanations. *Proc. R. Soc. London A*, 459:2821–2845, 2003.

[Moore, 1967] RE Moore. *Interval analysis.* Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

[Moore, 1979] RE Moore. *Methods and Applications of Interval analysis.* SIAM, Philadelphia, Pennsylvania, 1979.

[Mount, 1982] SM Mount. A catalogue of splice junction sequence. *Nucleic Acids Research*, 10(2):459–472, 1982.

[Myers, 1999] E Myers. Whole-genome dna sequencing. *IEEE Computational Engineering and Science*, 3(1):33–43, 1999.

[Nasrallah, 2002] JB Nasrallah. Recognition and rejection of self in plant reproduction. *Science*, 296:305–308, 2002.

[Neumaier, 1990] A Neumaier. *Interval methods for systems of equations.* Cambridge university press, 1990.

[Pachter and Sturmfels, 2004a] Lior Pachter and Bernd Sturmfels. Parametric inference for biological sequence analysis. *Proc Natl Acad Sci U S A*, 101(46):16138–43, 2004.

[Pachter and Sturmfels, 2004b] Lior Pachter and Bernd Sturmfels. Tropical geometry of statistical models. *Proc Natl Acad Sci U S A*, 101(46):16132–7, 2004.

[Pin, 1998] Jean-Eric Pin. Tropical semirings. In *Idempotency (Bristol, 1994)*, volume 11 of *Publ. Newton Inst.*, pages 50–69. Cambridge Univ. Press, Cambridge, 1998.

[Radmacher *et al.*, 2001] M.D. Radmacher, R. Simon, R. Desper, R. Taetle, A.A. Schäffer, and M.A. Nelson. Graph models of oncogenesis with an application to melanoma. *J. Theor. Biol.*, 212:535–548, 2001.

[Rall, 1981] LB Rall. *Automatic differentiation, techniques and applications*, volume 120 of *Springer lecture notes in computer science.* Springer-Verlag, New York, 1981.

[Ratz, 1992] D Ratz. *Automatische ergebnisverifikation bei globalen optimierungsproblemen.* Ph.D. dissertation, Universitat Karlsruhe, Karlsruhe, Germany, 1992.

[Sainudiin *et al.*, 2005] R Sainudiin, SW Wong, K Yogeeswaran, J Nasrallah, Z Yang, and R Nielsen. Detecting site-specific physicochemical selective pressures: applications to the class-I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. *Journal of Molecular Evolution*, in press, 2005.

[Sainudiin, 2004] R Sainudiin. Enclosing the maximum likelihood of the simplest DNA model evolving on fixed topologies: towards a rigorous framework for phylogenetic inference. Technical Report BU1653-M, Department of Biol. Stats. and Comp. Bio., Cornell University, 2004.

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

[Sandelin *et al.*, 2004] A. Sandelin, P. Bailey, S. Bruce, P.G. Engström, J.M. Klos, W.W. Wasserman, J. Ericson, and B. Lenhard. Arrays of ultraconserved noncoding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, 5:99, 2004.

[Sankoff and Blanchette, 2000] David Sankoff and Mathieu Blanchette. Comparative genomics via phylogenetic invariants for Jukes-Cantor semigroups. In *Stochastic models (Ottawa, ON, 1998)*, volume 26 of *CMS Conf. Proc.*, pages 399–418. Amer. Math. Soc., Providence, RI, 2000.

[Sankoff and Nadeau, 2003] D Sankoff and JH Nadeau. Chromosome rearrangements in evolution: from gene order to genome sequence and back. *Proceedings of the National Academy of Sciences, USA*, 100:11188–11189, 2003.

[Schrijver, 1986] Alexander Schrijver. *Theory of linear and integer programming*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons Ltd., Chichester, 1986. A Wiley-Interscience Publication.

[Schwartz *et al.*, 2003] S Schwartz, WJ Kent, A Smit, Z Zhang, R Baertsch, RC Hardison, D Haussler, and W Miller. Human-mouse alignments with blastz. *Genome Research*, 13:103–107, 2003.

[Semple and Steel, 2003] Charles Semple and Mike Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

[Simon *et al.*, 2000] R. Simon, R. Desper, C.H. Papadimitriou, A. Peng, D.S. Alberts, R. Taetle, J.M. Trent, and A.A. Schäffer. Chromosome abnormalities in ovarian adenocarcinoma: III. Using breakpoint data to infer and test mathematical models for oncogenesis. *Genes, Chromosomes & Cancer*, 28:106–120, 2000.

[Sneath and Sokal, 1973] Peter H.A. Sneath and Robert R. Sokal. *Numerical taxonomy: the principles and practice of numerical classification*. W.H. Freeman, San Francisco, 1973.

[Speyer and Sturmfels, 2004] David Speyer and Bernd Sturmfels. The tropical Grassmannian. *Adv. Geom.*, 4(3):389–411, 2004.

[Stanley, 1999] Richard P. Stanley. *Enumerative combinatorics. Vol. 2*, volume 62 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999. With a foreword by Gian-Carlo Rota and appendix 1 by Sergey Fomin.

[Steel *et al.*, 1992] M. A. Steel, M. D. Hendy, L. A. Székely, and P. L. Erdős. Spectral analysis and a closest tree method for genetic sequences. *Appl. Math. Lett.*, 5(6):63–67, 1992.

[Strassen, 1983] V. Strassen. Rank and optimal computation of generic tensors. *Linear Algebra Appl.*, 52/53:645–685, 1983.

[Strimmer and Moulton, 2000] K. Strimmer and V. Moulton. Likelihood analysis of phylogenetic networks using directed graphical models. *Molecular Biology and Evolution*, 17:875–881, 2000.

[Strimmer *et al.*, 2001] K. Strimmer, C. Wiuf, and V. Moulton. Recombination analysis using directed graphical models. *Molecular Biology and Evolution*, 18:97–99, 2001.

[Studier and Keppler, 1988] JA Studier and KJ Keppler. A note on the neighbor-joining method of saitou and nei. *Molecular Biology and Evolution*, 5:729–731, 1988.

[Sturmfels, 2002] Bernd Sturmfels. *Solving systems of polynomial equations*, volume 97 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC, 2002.

[Susko, 2003] E. Susko. Confidence regions and hypothesis tests for topologies using generalized least squares. *Molecular Biology and Evolution*, 2003.

[Székely *et al.*, 1993] L. A. Székely, M. A. Steel, and P. L. Erdős. Fourier calculus on evolutionary trees. *Adv. in Appl. Math.*, 14(2):200–210, 1993.

[Tesler, 2002] G Tesler. *Journal of Computer and System Sciences*, 65(3):587–609, 2002.

[Thomas *et al.*, 2003] J W Thomas, J W Touchman, R W Blakesley, G G Bouffard, S M Beckstrom-Sternberg, E H Margulies, M Blanchette, A C Siepel, P J Thomas, J C McDowell, B Maskeri, N F Hansen, M S Schwartz, R J Weber, W J Kent, D Karolchik, T C Bruen, R Bevan, D J Cutler, S Schwartz, L Elnitski, J R Idol, A B Prasad, S-Q Lee-Lin, V V B Maduro, T J Summers, M E Portnoy, N L Dietrich, N Akhter, K Ayele, B Benjamin, K Cariaga, C P Brinkley, S Y Brooks, S Granite, X Guan, J Gupta, P Haghighi, S-L Ho, M C Huang, E Karlins, P L Laric, R Legaspi, M J Lim, Q L Maduro, C A Masiello, S D Mastrian, J C McCloskey, R Pearson, S Stantripop, E E Tiongson, J T Tran, C Tsurgeon, J L Vogt, M A Walker, K D Wetherby, L S Wiggins, A C Young, L-H Zhang, K Osoegawa, B Zhu, B Zhao, C L Shu, P J De Jong, C E Lawrence, A F Smit, A Chakravarti, D Haussler, P Green, W Miller, and E D Green. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424(6950):788–93, 2003.

[Valiant, 1979] L Valiant. The complexity of computing the permanent. *Theoretical Computer Science*, 8:189–201, 1979.

[Varchenko, 1995] A. Varchenko. Critical points of the product of powers of linear functions and families of bases of singular vectors. *Compositio Math.*, 97(3):385–401, 1995.

[Venter *et al.*, 2001] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigo, M J Campbell, K V Sjolander, B Karlak, A Kejariwal,

H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. The sequence of the human genome. *Science*, 291(5507):1304–51, 2001.

[Viterbi, 1967] A.J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269, 1967.

[Vogelstein *et al.*, 1988] B. Vogelstein, E. Fearon, and S. Hamilton. Genetic alterations during colorectal-tumor development. *N. Engl. J. Med.*, 319:525–532, 1988.

[von Heydebreck *et al.*, 2004] A. von Heydebreck, B. Gunawan, and L. Füzesi. Maximum likelihood estimation of oncogenetic tree models. *Biostatistics*, 5(4):545–556, 2004.

[Warshall, 1962] S Warshall. A theorem on boolean matrices. *Journal of the ACM*, 9(1):18, 1962.

[Watson and Crick, 1953] J Watson and F Crick. A structure for deoxyribose nucleic acid. *Nature*, 171:964–967, 1953.

[Wetzel, 1995] R. Wetzel. Zur visualisierung abstrakter Ähnlichkeitsbeziehungen. Master's thesis, Fakultät Mathematik, Universität Bielefeld, 1995.

[Winkler, 1984] P. Winkler. Isometric embeddings in products of complete graphs. *Discrete Applied Mathematics*, 7:221–225, 1984.

[Wolf *et al.*, 2000] M. J. Wolf, S. Easteal, M. Kahn, B. D. McKay, and L. S. Jermiin. Trexml: A maximum likelihood program for extensive tree-space exploration. *Bioinformatics*, 16:383–394, 2000.

[Woolfe *et al.*, 2005] A. Woolfe, M. Goodson, D.K. Goode, P. Snell, G.K. McEwen, T. Vavouri, S.F. Smith, P. North, H. Callaway, K. Kelly, K. Walter, I. Abnizova, W. Gilks, Y.J.K. Edwards, J.E. Cooke, and G. Elgar. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biology*, 3:7, 2005.

[Wu and Li, 1985] Chung-I Wu and Wen-Hsiung Li. Evidence for Higher Rates of Nucleotide Substitution in Rodents Than in Man. *PNAS*, 82(6):1741–1745, 1985.

[Yap and Pachter, 2004] Von Bing Yap and Lior Pachter. Identification of evolutionary hotspots in the rodent genomes. *Genome Res*, 14(4):574–9, 2004.

[Yoder and Yang, 2000] A. D. Yoder and Z. Yang. Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution*, 17:1081–1090, 2000.

[Yu *et al.*, 2002] Jun Yu, Songnian Hu, Jun Wang, Gane Ka-Shu Wong, Songgang Li, Bin Liu, Yajun Deng, Li Dai, Yan Zhou, Xiuqing Zhang, Mengliang Cao, Jing Liu, Jiandong Sun, Jiabin Tang, Yanjiong Chen, Xiaobing Huang, Wei Lin, Chen Ye, Wei Tong, Lijuan Cong, Jianing Geng, Yujun Han, Lin Li, Wei Li, Guangqiang Hu, Xiangang Huang, Wenjie Li, Jian Li, Zhanwei Liu, Long Li, Jianping Liu, Qiuhui Qi, Jinsong Liu, Li Li, Tao Li, Xuegang Wang, Hong Lu, Tingting Wu, Miao Zhu, Peixiang Ni, Hua Han, Wei Dong, Xiaoyu Ren, Xiaoli Feng, Peng Cui, Xianran Li, Hao Wang, Xin Xu, Wenxue Zhai, Zhao Xu, Jinsong Zhang, Sijie He, Jianguo Zhang, Jichen Xu, Kunlin Zhang, Xianwu Zheng, Jianhai Dong, Wanyong

Zeng, Lin Tao, Jia Ye, Jun Tan, Xide Ren, Xuewei Chen, Jun He, Daofeng Liu, Wei Tian, Chaoguang Tian, Hongai Xia, Qiyu Bao, Gang Li, Hui Gao, Ting Cao, Juan Wang, Wenming Zhao, Ping Li, Wei Chen, Xudong Wang, Yong Zhang, Jianfei Hu, Jing Wang, Song Liu, Jian Yang, Guangyu Zhang, Yuqing Xiong, Zhijie Li, Long Mao, Chengshu Zhou, Zhen Zhu, Runsheng Chen, Bailin Hao, Weimou Zheng, Shouyi Chen, Wei Guo, Guojie Li, Siqi Liu, Ming Tao, Jian Wang, Lihuang Zhu, Longping Yuan, and Huanming Yang. A draft sequence of the rice genome (oryza sativa l. ssp. indica). *Science*, 296(5565):79–92, 2002.

[Yu *et al.*, 2005] J Yu, J Wang, W Lin, S Li, H Li, J Zhou, P Ni, W Dong, S Hu, C Zeng, J Zhang, Y Zhang, R Li, Z Xu, X Li, H Zheng, L Cong, L Lin, J Yin, J Geng, G Li, J Shi, J Liu, H Lv, J Li, Y Deng, L Ran, X Shi, X Wang, Q Wu, C Li, X Ren, D Li, D Liu, X Zhang, Z Ji, W Zhao, Y Sun, Z Zhang, J Bao, Y Han, L Dong, J Ji, P Chen, S Wu, Y Xiao, D Bu, J Tan, L Yang, C Ye, J Xu, Y Zhou, Y Yu, B Zhang, S Zhuang, H Wei, B Liu, M Lei, H Yu, Y Li, H Xu, S Wei, X He, L Fang, X Huang, Z Su, W Tong, Z Tong, J Ye, L Wang, T Lei, C Chen, H Chen, H Huang, F Zhang, N Li, C Zhao, Y Huang, L Li, Y Xi, Q Qi, W Li, W Hu, X Tian, Y Jiao, X Liang, J Jin, L Gao, W Zheng, B Hao, S Liu, W Wang, L Yuan, M Cao, J McDermott, R Samudrala, GK Wong, and H Yang. The genomes of oryza sativa: A history of duplications., 2005.

[Z. and D., 1995] Yang Z. and Roberts. D. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Molecular Biology and Evolution*, 12:451–458, 1995.

[Zang, 2001] K.D. Zang. Meningioma: a cytogenetic model of a complex benign human tumor, including data on 394 karyotyped cases. *Cytogenet. Cell Genet.*, 93:207–220, 2001.

[Ziegler, 1995] Günter M. Ziegler. *Lectures on polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995.