

## Lecture 1 — January 22

Lecturer: Lior Pachter

Scribe/ Editor: Robert Bradley/ Cordelia Csar

## 1.1 Course requirements

- Scribe and edit notes
- Homework (+/- grading); research questions intermixed. Homework is due approximately one week after it is assigned.
- Final project

This lecture is intended as an overview of some of the objects we will encounter in the first part of the course. A key idea is that of a *laminar family*— a concept that originates with Darwin.

## 1.2 Laminar families and monophyletic groups

**Definition 1.1.** A set system  $\mathcal{H}$  is a collection of subsets of a set  $X$ .  $\mathcal{H}$  is also called a hypergraph. The elements of  $X$  are called vertices. The elements of  $\mathcal{H}$  are called edges.

**Definition 1.2.** A laminar family or hierarchical system is a hypergraph  $\mathcal{H}$  that has the following property: For any two edges  $e_1, e_2 \in \mathcal{H}$ , either  $e_1 \cap e_2 = \emptyset$  or  $e_1 \subseteq e_2$  or  $e_2 \subseteq e_1$ .

**Definition 1.3.** A monophyletic group of species is an ancestor together with **all** of the descendant species. We generally speak of a monophyletic group as a set of extant species together with their MRCA (most recent common ancestor).

**Definition 1.4.** A paraphyletic group of species is an ancestor together with **some** of the descendant species.

Monophyletic groups (of nucleotides) form laminar families. We term these *alignments*.

**Example.**  $\{\text{human}\}, \{\text{chimp}\}, \{\text{human, chimp}\}$  is a monophyletic group. Here the elements of  $X$  are species and the edges of  $\mathcal{H}$  are sets of species representing ancestors.

**Example.**  $\{\text{human}\}, \{\text{chimp}\}, \{\text{rat}\}, \{\text{human, chimp, rat}\}$  is not a monophyletic group.

### 1.3 Laminar families and partially ordered sets

**Definition 1.5.** A partially ordered set (POSET) is a set  $P$  together with a relation  $\leq_P$  satisfying:

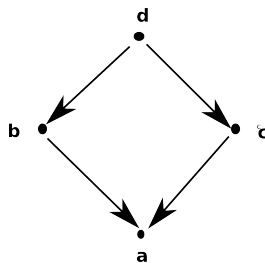
1.  $x \leq_P x \forall x \in P$
2. if  $x \leq_P y$  and  $y \leq_P x$ , then  $x = y \forall x, y \in P$
3. if  $x \leq_P y$  and  $y \leq_P z$ , then  $x \leq_P z \forall x, y \in P$

Laminar families give rise to partially ordered sets (POSETs): We can impose a relation based on ancestry:  $e_1 \leq e_2$  iff  $e_1 \subseteq e_2$ .

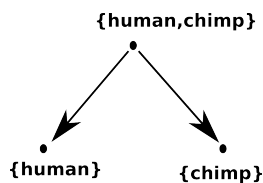
We can represent a POSET with a “Hasse diagram,” a directed acyclic graph (DAG) where the vertices are elements of the set  $P$  and there exists an edge  $x \rightarrow y$  iff  $y \leq_P x$  and  $\nexists z$  such that  $x \leq_P z \leq_P y$ .

**Example.** Hasse diagram for the POSET

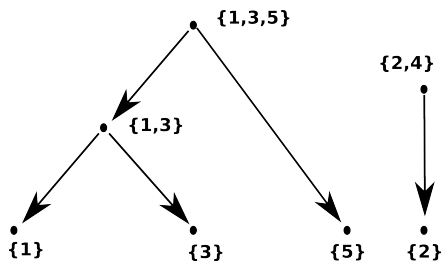
$$\begin{aligned} a &\leq b & b &\leq d \\ c &\leq d & a &\leq c \\ a &\leq d & & \end{aligned}$$



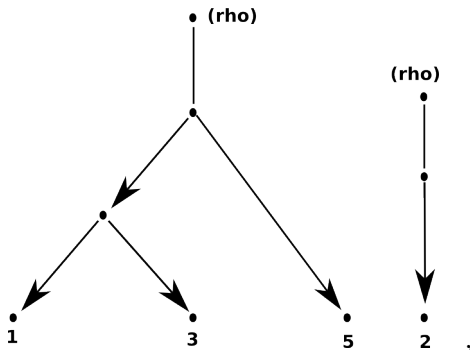
**Example.** Human and chimp.



**Example.**  $\mathcal{H} = \{\{2, 4\}, \{1\}, \{3\}, \{1, 3\}, \{5\}, \{1, 3, 5\}, \{2\}\}$



The leaves in this diagram are “atoms” of the POSET. Note that this gives us a partial labeling of nodes,



where we have rooted the trees.

**Definition 1.6.** A root of a tree is a specially designated leaf labeled  $\rho$ .

### 1.4 Laminar families and rooted $X$ -forests

**Definition 1.7.** Let  $X$  be a set. A rooted  $X$ -forest  $F$  is:

1. A collection of rooted trees.
2. A function  $\phi : X \rightarrow V(F)$  providing a partial labeling of (non-root) vertices such that every unlabeled vertex has degree  $> 2$ .

**Theorem 1.8.** (Edmonds-Giles 1977)

There is a bijection between laminar families for  $X$  and rooted  $X$ -forests.

**Proof:** Given a rooted  $X$ -forest  $F$ , we build the edges  $e$  of a laminar family  $\mathcal{H}$  in the following way. For every vertex  $v$  in  $F$ , we do the following:  
 Removing the edge above  $v$  yields two components of the tree containing  $v$ , one component

containing the root  $\rho$  and one not. We get a corresponding edge  $e \in \mathcal{H}$  where  $x \in e$  for all  $x \in X$  in the component not containing the root.

We need to check that either  $e_1 \cap e_2 = \emptyset$  or  $e_1 \subseteq e_2$  or  $e_2 \subseteq e_1$ . This is easily seen: there exists a unique path between any 2 vertices of a tree, so our construction of the edges of  $\mathcal{H}$  satisfies this condition.

To complete the proof we need to go in the other direction as well (give a construction for a rooted  $X$ -forest from a laminar family). This is assigned as homework.  $\square$

Some notation and technicalities:

1. For laminar families, we require  $\forall x \in X, x \in e$  for some  $e \in \mathcal{H}$ . We refer to these as “complete laminar families.”
2. Let  $\mathcal{L}_N$  denote all complete laminar families on a set  $X$  with  $|X| = N$ . Note that there is a partial order defined on  $\mathcal{L}_N$ :

$$\mathcal{H}_1, \mathcal{H}_2 \in \mathcal{L}_N$$

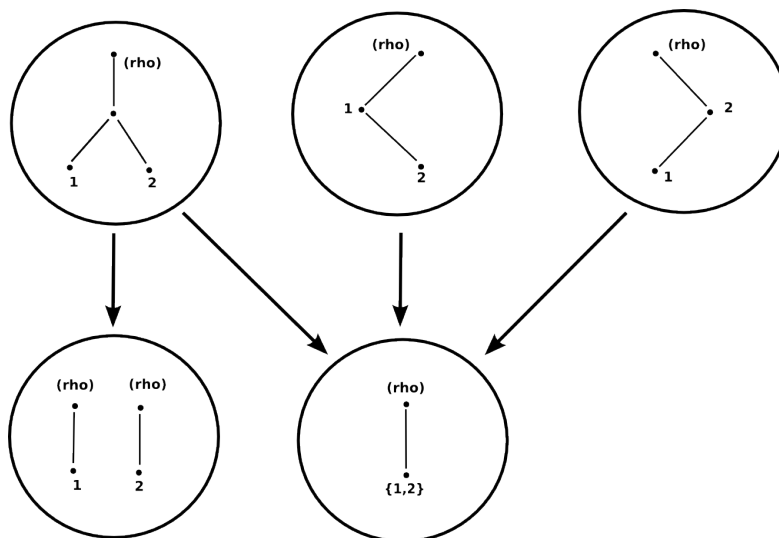
$$\mathcal{H}_1 \leq \mathcal{H}_2 \quad \text{iff} \quad \forall e \in \mathcal{H}_1, e \in \mathcal{H}_2 \tag{1.1}$$

3. A “resolved laminar family” is such that all internal vertices of the corresponding  $X$ -forest have degree  $\geq 3$ . An upper bound on the number of resolved complete laminar families is

$$\sum_k \frac{(2N - k - 1)! 2^k}{(N - k)! (k - 1)!} \tag{1.2}$$

4. An “extant laminar family” is such that  $\forall x \in X, \{x\} \in \mathcal{H}$ .

**Example.** Consider  $\mathcal{L}_2$ , where  $X = \{1, 2\}$ . We can use (1.1) to obtain a partial ordering for  $\mathcal{L}_2$ . The corresponding Hasse diagram is:



Each circle represents one of the possible complete laminar families. The arrows are edges of the Hasse diagram. Our main objective in the first part of the course will be to explain the connection between this poset and the *space of evolutionary models*. In particular, we will encounter a closely related poset, the *Tuffley poset*, whose geometric realization is the space of evolutionary models. Evolutionary models are special types of *discrete statistical models*. Next, we briefly explain our approach to such models.

## 1.5 Statistical models

In the first part of the course we will study statistical models associated with laminar families. By this we mean the following:

**Definition 1.9.** A statistical model is a family of probability distributions on a finite set  $\{1, \dots, m\}$ ,

$$(p_1, \dots, p_m) \in \mathbb{R}^m : \sum_{i=1}^m p_i = 1 \text{ and } p_j \geq 0 \forall j.$$

For  $\Sigma$  a finite set (e.g. the nucleotides), we might have statistical models over  $\Sigma^{|X|}$ ,  $\Sigma^{|\mathcal{H}|}$  or  $\mathcal{L}_N$ .

We will look at parametric models,

$$\mathbb{R}^d \rightarrow \mathbb{R}^m,$$

where  $d$  = number of parameters of the model, and particularly at log-linear (toric) models.

**Example.** Independence model over  $\{00, 01, 10, 11\}$  ( $m = 4$ ). The model is

$$f : \mathbb{R}^4 \rightarrow \mathbb{R}^4 \tag{1.3}$$

$$(\theta_1, \theta_2, \theta_3, \theta_4) \rightarrow (\theta_1\theta_3, \theta_1\theta_4, \theta_2\theta_3, \theta_2\theta_4), \tag{1.4}$$

where the parameters  $\theta_{1,2,3,4}$  ( $\theta_1 = P$  (first character is 0), etc.) satisfy  $\theta_1 + \theta_2 = \theta_3 + \theta_4 = 1$ .

Note that  $p_{00}p_{11} = p_{10}p_{01}$  for this model, and furthermore that this relation defines a surface, the “**Segre variety**,” in a tetrahedron. In this course we will frequently adopt the geometric point of view.

**Example. Simpson’s paradox.** We are interested in knowing how a particular allele correlates with disease. We have statistics for both Caucasian and Asian groups, shown left-to-right here:

$$\begin{pmatrix} 7 & 30 \\ 2 & 15 \end{pmatrix} \quad \begin{pmatrix} 6 & 3 \\ 15 & 17 \end{pmatrix}$$

where the columns are (sick, healthy) and the rows are (I, II). We calculate the *odds-ratios*

$$\frac{u_{00}u_{11}}{u_{10}u_{01}},$$

where the  $u_{ij}$  are the number of observations of each type. If the categories are independent, we expect the empirical distribution to lie on the model, which means that  $u_{00}u_{11} = u_{10}u_{01}$ . In the example, the odds-ratios are

$$\frac{7 * 15}{2 * 30} = 1.75 \qquad \frac{6 * 17}{15 * 3} = 2.26.$$

We could alternately consider the combined statistics for both Caucasians and Asians,

$$\begin{pmatrix} 13 & 33 \\ 17 & 32 \end{pmatrix},$$

giving an odds-ratio of

$$\frac{13 * 32}{17 * 33} = 0.74.$$

This is known as *Simpson's paradox*. The corresponding geometric statement is one about convexity. In fact, the individual odds ratios lie on one side of the independence model, and the combined odds ratio is their midpoint: a point in the tetrahedron on the opposite side.

## 1.6 Homework

**Proof of Edmonds-Giles theorem.** Complete Proof 1.8.

**Enumeration of resolved complete laminar families.** Derive the upper bound (1.2) on the number of resolved complete laminar families.

**Enumeration of complete laminar families.** Count the number of complete laminar families. (This is an open problem!)