On the comparison of incompatibility of split systems across different taxa sizes

Michael Hendriksen, Nils Kapust

Institut für Molekular Evolution, Heinrich-Heine Universität
April 2, 2020

Abstract

The concept of k-compatibility measures how many phylogenetic trees it would take to display all splits in a given set. A set of trees that display every single possible split is termed a *universal tree set*. In this note, we find $A(n)$, the minimal size of a universal tree set for n taxa. By normalising the k-compatibility using $A(n)$, one can then compare incompatibility of split systems across different taxa sizes. We demonstrate this application by comparing two SplitsTree networks of different sizes derived from archaeal genomes.

Keywords— phylogenetic trees, compatibility, split systems, bipartitions, matchings, SplitsTree, archaeal genomes

1 Introduction

Phylogenetic trees are ubiquitously used to represent the evolutionary history of organisms [3]. An individual tree is equivalent to a set of splits, which each indicate that for the sets of organisms A and B, organisms are more closely related within the set than between them. However, data can often produce conflicting results, whether through measurement error or complex biological phenomena such as incomplete lineage sorting or horizontal gene transfer. This can result in splits that contradict each other.

This naturally gave rise to the definition of k-compatibility, first studied in [2], which, loosely, measures how many phylogenetic trees it would take to display all splits indicated by the data. We say that a set of trees that display all splits is a *universal tree set*.

In the present paper, we consider the question of maximum possible incompatibility - that is, given a set of taxa X of size n, how large is a universal tree set of minimal size?

This function can also be characterised as finding minimal k such that every split system S on X is k-compatible, or finding the minimal k so that the split system consisting of every split is k-compatibile.

By characterising $A(n)$, one can then contextualise a split system in terms of how incompatible it is compared to the worst case scenario. Further, by normalising the k-compatibility using $A(n)$, we can then compare incompatibility of split systems across different taxa sizes.

Of particular interest is the fact that the widely-used SplitsTree software [4] creates a so-called *split network*, which is used to represent conflicting split signals from data. The present results will now allow those who use SplitsTree to fairly compare incompatibility of data across different taxa sizes.

In Section 2 we provide background information. In Section 3 we prove some lemmata on bipartitions. In Section 4 we apply these lemmata and some classical theorems to prove the main result.

2 Background

A phylogenetic tree on a set of taxa X is an acyclic graph (V, E) such that there are no vertices of degree-2 and the degree-1 vertices (termed *leaves*) are bijectively labelled by the elements of X.

Recall that a *split* $A|B$ of a set X is a bipartition of X into two sets A, B; where $B = X \backslash A$. Define the *size* of a split $A|B$ to be $\min(|A|, |B|)$. We denote by $\mathcal{S}(X)$ the set of all splits of X, and any subset S of $\mathcal{S}(X)$ is called a *split system* on X.

Given a phylogenetic tree $T = (V, E)$ on X, each edge can be associated with a split in the following way. If an edge e is deleted from T, this disconnects the graph into two components, each with at least one labelled vertex. This naturally induces a bipartition on the leaf set $A|B$, which we call the split *associated* with e, and we say that $A|B$ is *displayed* by T if it is associated with some edge of T.

It is well-known [9] that two splits $A|B$ and $C|D$ can only be displayed by the same phylogenetic tree if one of the four intersections

$$A \cap C, A \cap D, B \cap C, B \cap D$$

is empty. If this condition is met by each pair of splits in a split system S, we say that S is *pairwise compatible*.

In a biological context, sets of incompatible splits frequently arise from data, and biologists wish to quantify the extent to which the set is incompatible. This naturally gave rise to the definition of k-compatibility. We say that a split system is *k-compatible* if it does not contain a subset of $k + 1$ pairwise incompatible splits. Equivalently, a split system S is k-compatible if one can find a set of k trees $\mathcal{T}$ such that every split in S is displayed by at least one tree in $\mathcal{T}$.

We say that a set of k trees $\mathcal{T}$ that displays every split in a set of splits S is *minimal* with respect to S if there are no sets of $k - 1$ trees with this property. If $\mathcal{T}$ contains trees that display every split in $\mathcal{S}(X)$, we say that $\mathcal{T}$ is a *universal tree set*. An example of a minimal universal tree set for 5 leaves is shown in Figure 1. Define the function $A(n)$ to be the value of $|\mathcal{T}|$, where $\mathcal{T}$ is a minimal universal tree set on a set of taxa of size n.

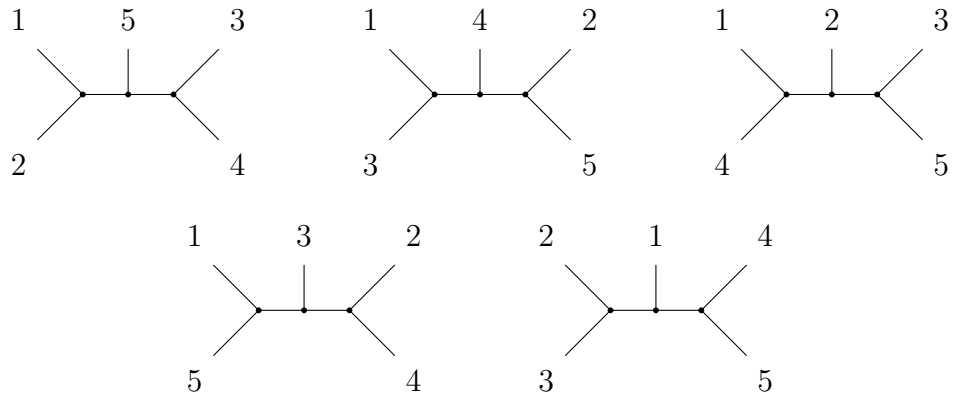

Figure 1: A minimal universal tree set on 5 leaves.

3 Combinatorial Results on Bipartitions

We will first derive some combinatorial lemmata on bipartitions that will assist with the main result.

Lemma 3.1. *Let n be an even integer. Then a tree on n leaves has at most one split of size $n/2$.*

Proof. Let $A|B$ and $C|D$ be a pair of splits on the same tree of size $n/2$. Then one of $A \cap C, A \cap D, B \cap C$ or $B \cap D$ is empty. Without loss of generality, suppose that $A \cap C$ is empty. Then $A \subseteq (X \backslash C) = D$, but since both partitions are of size $n/2$, it follows that $A = D$, so $B = C$ and thus $A|B$ and $C|D$ are equivalent partitions. The lemma follows. □

Note that of course a tree need not have any such split, as we can consider the star tree for any number of leaves $n \geq 4$. This Lemma gives a lower bound for A(n) for even n, as trees with an even number of leaves can have at most one split of size $\frac{n}{2}$, and $\frac{1}{2}\binom{n}{\frac{n}{2}}$ is the number of such splits for a given n. In fact, $A(n)$ actually equals this lower bound in the even case, as we will see in Theorem 4.3.

Lemma 3.2. *Let $n = 2m + 1$ where $m \geq 2$ is a positive integer. Then a tree on n leaves displays at most two splits of size m.*

Proof. Seeking a contradiction, let $A|B$, $C|D$ and $E|F$ be three distinct splits on the same tree, so that $|A| = |C| = |E| = m, |B| = |D| = |F| = m+1$. Then one of $A \cap C, A \cap D, B \cap C$ or $B \cap D$ is empty. As $A|B$ and $C|D$ are distinct splits, it must be the case that $A \cap C$ is empty. Therefore, $A \subset D$ and $C \subset B$; in fact, $D = A \cup \{x\}$ for some taxon x. This implies that $C = B \backslash \{x\}$.

By similar logic, $F = A \cup \{y\}$ and $E = B \backslash \{y\}$ for some taxon y so that $y \neq x$ (since $E|F$ and $C|D$ are distinct). But then $C \cap E, C \cap F, D \cap E$ and $D \cap F$ must all be non-empty, which is a contradiction. The lemma follows. □

One can observe that in the $n = 3$ case there are 3 such splits (and that a minimal universal tree set on 3 leaves consists of just the star tree on 3 leaves). Outside of this case, as each tree with an odd number of leaves can display up to two of these splits, a natural

follow-up question is whether there are any obstructions to pairing all such splits in this way. Fortunately we can, using the concept of matchings. We will need two definitions before we can see this.

Definition 3.3. A *matching* M of a graph G is a set of edges of G such that no two edges share the same vertex. A *defect-d* matching M is a matching so that all except d vertices of G have an incident edge from M. Defect-0 matchings are also referred to as *perfect* matchings.

Definition 3.4. A graph G is said to be *vertex-transitive* if, given any two vertices v_1 and v_2 of G, there is some automorphism

$$f\colon V(G) \to V(G)$$

such that

$$f(v_1) = v_2, f(v_1) = v_2.$$

We can now state the following theorem from [6].

Theorem 3.5. *Every connected vertex-transitive graph with an even number of vertices has a perfect matching and every connected vertex-transitive graph with an odd number of vertices has a defect-*1 *matching.*

We will construct a graph related to the compatibility of bipartitions, which may have either an odd or even number of vertices. In order to distinguish between these cases we will also need the following theorem.

Theorem 3.6 (Kummer's Theorem [5])**.** *If p is a prime, then the largest power of p that divides $\binom{m+n}{n}$, for m and n non-negative, is the number of carries when m and n are added in base q.*

As a simple corollary to this, we get

Corollary 3.7. *If $n = 2m + 1$ for positive integer m, then $\binom{n}{m}$ is odd if and only if m is a power of* 2*.*

Let $Bip(n, m)$ be the set of bipartitions of n of size m.

Lemma 3.8. *Let $n = 2m + 1$, where $m \geq 2$. If $m = 2^k$ for some integer k then we can partition $Bip(n, m)$ into compatible pairs $\{A|B, C|D\}$with one left over, otherwise we can partition $Bip(n, m)$ into compatible pairs with no leftovers.*

Proof. Let G be the graph whose vertices are the elements of $Bip(n, m)$, and there is an edge between vertices $A|B$ and $C|D$ if and only if they are compatible bipartitions. This graph is certainly connected, and we will show that it is vertex-transitive.

Let $A|B$ and $C|D$ be two vertices of G, where $|A| = |C| = m, |B| = |D| = m + 1$. Let σ be a permutation of X so that σ applied to each taxon in A obtains C, and similarly applied to B obtains D. Then the induced action by applying this permutation to every bipartition in G is an automorphism that maps $A|B$ to $C|D$, hence G is vertex-transitive.

We now apply Theorem 3.5 and Kummer's Theorem and the lemma is immediately proven. □

4 Minimal Universal Tree Sets

We are now, barring the statement of a few useful classical theorems from extremal set theory, ready to prove the main result. Define $P(X)$ to be the poset on the power set of X ordered by set inclusion.

Theorem 4.1 (Sperner's Theorem [8])**.** *Let X be a set of size n. Then the largest antichain in $P(X)$ has size $\binom{n}{\lfloor n/2 \rfloor}$*

Theorem 4.2 (Dilworth's Theorem [1])**.** *Let P be a poset and suppose the largest antichain in P has size r. Then P can be partitioned into r chains.*

Theorem 4.3. *Let X be a set of size n. Then a minimal universal tree set for X has size*

$$A(n) = \left\lceil \frac{1}{2} \binom{n}{\lfloor \frac{n}{2} \rfloor} \right\rceil.$$

Proof. Consider the poset $P(X)$. Sperner's Theorem states that the largest antichain of $P(X)$ has size $\binom{n}{\lfloor n/2 \rfloor}$, and Dilworth's Theorem implies that we can therefore partition $P(X)$ into $\binom{n}{\lfloor n/2 \rfloor}$ chains, each of which contains one subset of size $n/2$ if n is even, or one subset of size $(n-1)/2$ if n is odd.

If n is even (resp. odd), consider the sub-poset $Chain(X)$ consisting of those non-empty subsets of size $n/2$ or less (resp. $(n-1)/2$ or less) $P'(X)$ and only those edges present in the chains. Let

$$\gamma : V(Chain(X)) \to \mathcal{S}(X)$$

be the function that maps the subset A to the bipartition $A|(X \backslash A)$. We apply the function γ to the vertices of $Chain(X)$ and identify any vertices $A|(X \backslash A)$ and $B|(X \backslash B)$ that now have the same label, which will only occur if n is even and $A = (X \backslash B)$.

If $n = 2m + 1$ is odd, we perform one additional modification. By Theorem 3.5, there exists a matching M between the bipartitions of the form $A|B$ where $|A| = m, |B| = m+1$ with at most one unpaired bipartition. For each such matching $(A|B, C|D)$ in M, add this edge to $\gamma(Chain(X))$.

Now, in either case, the number of components of $\gamma(Chain(X))$ will be

$$k = \left\lceil \frac{1}{2} \binom{n}{\lfloor \frac{n}{2} \rfloor} \right\rceil.$$

We construct a universal tree set as follows. Denote the components of $\gamma(Chain(X))$ by $C_1, ..., C_k$.

We claim that the set $V(C_1), ..., V(C_k)$ is a universal tree set, in particular that all $V(C_i)$ are sets of compatible bipartitions.

First suppose n is even and let the unique bipartition of size $n/2$ in $V(C_i)$ be $A|B$. Suppose we have two distinct bipartitions, $C|D$ and $C'|D'$. In order to show compatibility, it suffices (but is not necessary) to show that one of C or D is contained in one of C' or D', or the reverse as the inclusion requirement quickly holds.

If $C'|D' = A|B$, then certainly C or D is a subset of A or B by the chain construction. We therefore assume neither bipartition is $A|B$, and without loss of generality that $|C| <$

$|D|$ and $|C'| < |D'|$. Then, due to the chain construction, either C and C' are subpartitions of the same partition (A or B), or one is a subset of A and the other of B. If C and C' are subpartitions of the same partition, then C and C' are from the same chain and therefore $C \subseteq C'$ or the reverse. However, if they are subpartitions of different partitions, then $C \subset X \backslash C' = D'$. Therefore in all cases $V(C_i)$ is a compatible tree set, and as it contains all possible bipartitions it is universal.

If n is odd, the options proceed analogously if $C|D$ and $C'|D'$ were obtained from the same chain in $P'(X)$. If not, then let the two (distinct) splits into partitions of size m be $A|B$ and $A'|B'$, so that $C|D$ was obtained from the same chain as $A|B$ and $C'|D'$ the same chain as $A'|B'$. In particular, without loss of generality suppose that $C \subseteq A$ and $C' \subseteq A'$.

As $A|B$ and $A'|B'$ are compatible, one of $A \cap A', A \cap B', B \cap A'$ or $B \cap B'$ are empty, but as B and B' have size $m+1$ the only possibility is that $A \cap A'$ is empty. Then as $C \subseteq A$ and $C' \subseteq A'$ it follows that $C \cap C'$ is empty, so $C|D$ and $C'|D'$ are compatible.

It finally remains to confirm that

$$k = \left\lceil \frac{1}{2} \binom{n}{\lfloor \frac{n}{2} \rfloor} \right\rceil .$$

is the minimal possible value. In the even case, as each tree in any universal tree set can contain at most one split of size $n/2$ by Lemma 3.1 (of which there are $\binom{n}{n/2}/2$, which is equal to the desired formula when n is even), the set is minimal.

In the odd case $n = 2m+1$, by Lemma 3.2 there can be at most two splits with a partition of size m, of which there are $\binom{n}{\lfloor n/2 \rfloor}$, so we obtain the desired formula in this case too (noting that in the odd case the binomial coefficient can be odd, hence the ceiling function). □

We note here that the associated integer sequence: $1, 1, 1, 3, 5, 10, 18, 35, 63...$ does not appear in the OEIS, and is in the process of submission.

There are several natural extensions to this problem for future research. For instance, one avenue could be to investigate how the value changes if we instead ask for a minimal universal set of networks with at most k reticulations. The probabilistic analogue of the question would also be interesting - how likely is it, given a set of k trees, to have a universal tree set (in particular for the minimal case, $k = A(n)$)?

5 Applications

We now consider two SplitsTree [4] networks derived from archaeal genomes and depicted in Figure 2. The first network N_1 contains 13 taxa, and the second network, N_2 is the network obtained after removal of the single taxon *Methanococcus maripaludis*, leaving 12 taxa.

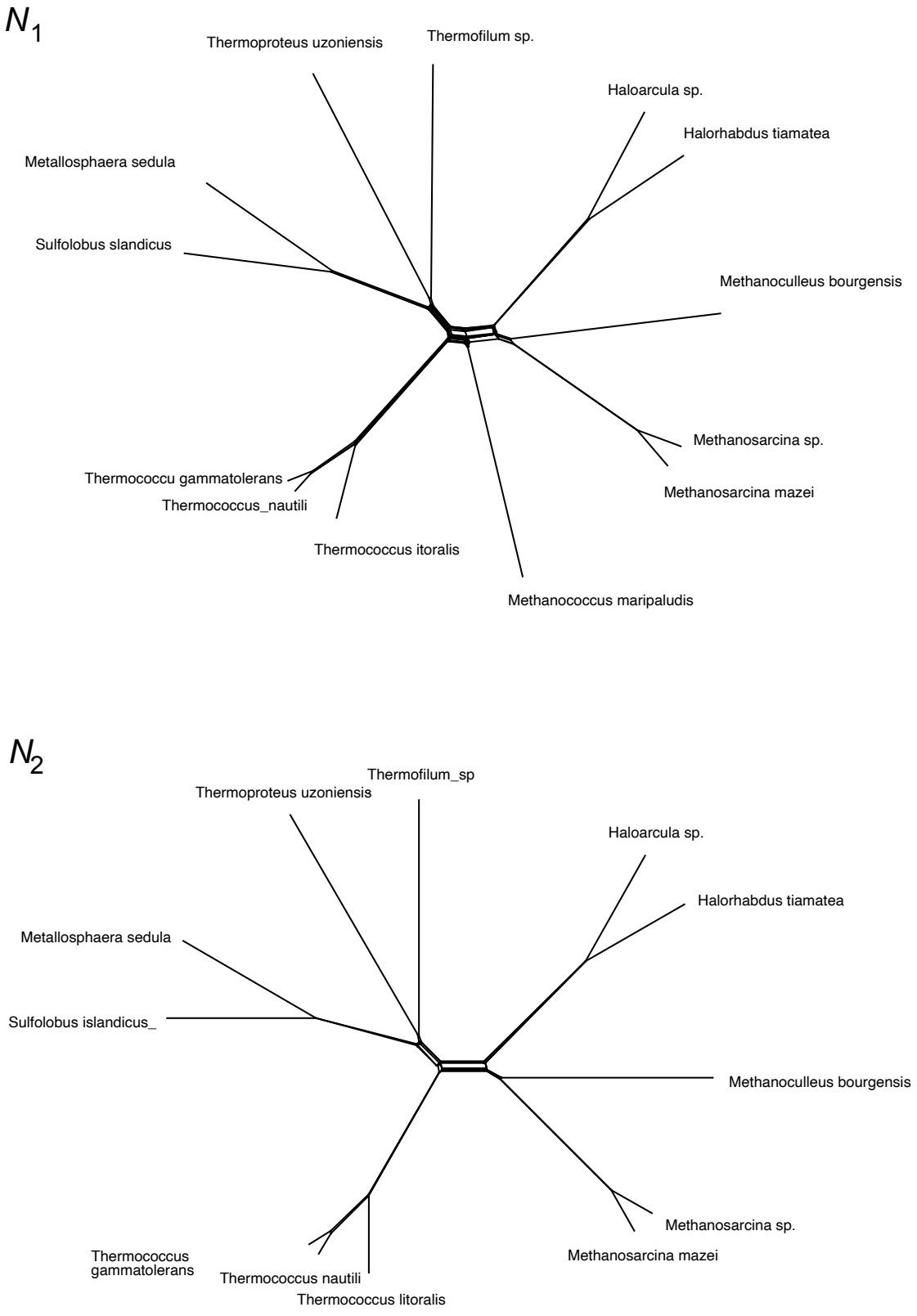

Figure 2: Networks N_1 and N_2 derived from archaeal genomes.

To generate these networks, 39 universal archaeal proteins gathered from Nelson-Sathi et al. [7], were used for a BLAST search against archaeal genomes obtained from the RefSeq 2016 database with an identity threshold of 20% and an e-value cut-off of 10^{-5}. For each, best hit alignments were generated and concatenated. These concatenated alignments were used to draw a Neighbor-Net using SplitsTree4.

Using a short Python program, we analysed the splits corresponding to each network and found a set of 4 trees that display all splits in N_1, showing 4-compatibility, and a set of 3 trees that display all splits in N_2, showing 3-compatibility. These were then shown to be minimal by hand - both networks display 3 incompatible splits of size 6, and N_1 additionally displays a partition $A|B$ of size 5 that is incompatible with each of the first 3. Therefore, if we denote the k-compatibility of the split system associated with a network N by $\kappa(N)$, we know $\kappa(N_1) = 4$ and $\kappa(N_2) = 3$.

If we denote the number of leaves of a network N by $|N|$, then we can define the *normalized k-compatibility* $Norm(N)$ to be

$$Norm(N) = \frac{\kappa(N)}{|N|}.$$

Now, as $A_T(13) = 858$ and $A_T(12) = 462$, we can normalize these k-compatibilities, and as

$$Norm(N_1) = \frac{4}{858} < Norm(N_2) = \frac{3}{462},$$

from the perspective relative to maximum incompatibility, N_2 is 'more incompatible' than N_1.

Acknowledgements

The authors thanks Prof. Dr. W. F. Martin, the Volkswagen Foundation 93_046 grant and the ERC Advanced Grant No. 666053 for their support during this research. The authors would also like to thank Andrew Francis for helpful comments on a draft, and Mike Steel, Fernando Tria and Falk Nagies for illuminating conversations on this topic.

Author Contributions

MH designed the paper, performed all mathematical research and wrote the majority of the paper. NK prepared and analysed the data, and wrote the Applications section.

References

[1] R. P. Dilworth. A decomposition theorem for partially ordered sets. (1950).

[2] A. Dress et al. $2kn - \binom{2k+1}{2}$: A Note on Extremal Combinatorics of Cyclic Split Systems. *Séminaire Lotharingien de Combinatoire*, 47 (2001).

[3] J. Felsenstein. *Inferring phylogenies*. Vol. 2. Sinauer associates Sunderland, MA, 2004.

[4] D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution* 23.2 (2006), pp. 254267.

[5] E. E. Kummer. Über die Ergänzungssätze zu den allgemeinen Reciprocitätsgesetzen. *Journal für die reine und angewandte Mathematik* 44 (1852), pp. 93146.

[6] C. H. C. Little, D. D. Grant, and D. A. Holton. On defect-d matchings in graphs. *Discrete Mathematics* 13.1 (1975), pp. 4154.

[7] S. Nelson-Sathi *et al.*. Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517.7532 (2015), pp. 7780.

[8] E. Sperner. Ein satz über untermengen einer endlichen menge. *Mathematische Zeitschrift* 27.1 (1928), pp. 544548.

[9] M. Steel. *Phylogeny: discrete and random processes in evolution*. SIAM, 2016.