

# Notes 2 : Splits-Equivalence Theorem

MATH 833 - Fall 2012

Lecturer: Sebastien Roch

References: [SS03, Chapters 3, 4]

## 1 Characters

Biological data can be formalized using the following definition.

**DEF 2.1 (Characters)** Let  $C$  be a set of character states. A (full) character on  $X$  is a function from  $X$  to  $C$ . A character is binary if  $|C| = 2$ .

**DEF 2.2 (Character Convexity)** A character  $\chi$  is convex on an  $X$ -tree  $\mathcal{T} = (T, \phi)$  with  $T = (V, E)$  if there is a function  $\bar{\chi} : V \rightarrow C$  such that

1.  $\bar{\chi} \circ \phi = \chi$  (i.e.,  $\bar{\chi}$  is an extension of  $\chi$  to all vertices).
2. For each  $\alpha \in C$  the subgraph of  $T$  induced by  $\{v \in V : \bar{\chi}(v) = \alpha\}$  is connected (i.e., any particular state transition (or its reverse) occurs only once in the tree).

Character convexity corresponds to evolutionary innovations occurring only once in the tree of life, that is, in the absence of *reverse transition* (a new state arising but later reverting to its earlier state) and *convergent transition* (a new state occurring in two different parts of the tree).

**DEF 2.3 (Character compatibility)** A collection of characters on  $X$  is compatible if there is an  $X$ -tree on which all of them are convex.

Finding such a tree is known as the *perfect phylogeny problem* and, in the binary character case, serves as a motivation (among others) to derive the Splits-Equivalence Theorem which we now consider.

## 2 Statement of the Splits-Equivalence Theorem

**DEF 2.4 (*X*-splits)** An *X*-split  $A|B$  is a nontrivial bipartition of  $X$ . Let  $\mathcal{T} = (T, \phi)$  be an *X*-tree with  $T = (V, E)$ . To each edge  $e$  of  $T$  corresponds an *X*-split as follows:  $T \setminus e$  consists of two components with vertex sets  $V_1, V_2$ ;  $\phi^{-1}(V_1)|\phi^{-1}(V_2)$  is the *X*-split corresponding to  $e$ . We denote by  $\Sigma(\mathcal{T})$  the collection of splits induced by  $\mathcal{T}$ .

**DEF 2.5 (Split Compatibility)** *X*-splits  $A_1|B_2$  and  $A_2|B_2$  are compatible if at least one of the sets  $C_1 = A_1 \cap A_2$ ,  $C_2 = A_1 \cap B_2$ ,  $C_3 = B_1 \cap A_2$  and  $C_4 = B_1 \cap B_2$  is empty. (Any two  $C_i$ s corresponding to a partition of  $A_1$ ,  $B_1$ ,  $A_2$ , or  $B_2$  must have a non-empty union. In particular, at least two of the  $C_i$ s must be non-empty and equality happens exactly when the splits are identical.)

It is straightforward to check that the splits induced by an *X*-tree are pairwise compatible. There is also a converse.

**THM 2.6 (Splits-Equivalence Theorem)** Let  $\Sigma$  be a collection of *X*-splits. Then,  $\Sigma = \Sigma(\mathcal{T})$  for some *X*-tree  $\mathcal{T}$  if and only if the splits in  $\Sigma$  are pairwise compatible. Such tree is unique up to isomorphism.

The easy direction follows from the following lemma.

**LEM 2.7** Let  $\sigma_1 \neq \sigma_2 \in \Sigma(\mathcal{T})$ . Then  $X$  can be partitioned into three sets  $X_1, X_2, X_3$  such that  $\sigma_1 = X_1|(X_2 \cup X_3)$ ,  $\sigma_2 = (X_1 \cup X_2)|X_3$  and  $X_1 \cap X_3 = \emptyset$ .

**Proof:** By definition,  $\sigma_i$  corresponds to an edge  $e_i = \{u_i, v_i\}$  of  $\mathcal{T}$ . W.l.o.g., there is a path connecting  $e_1$  and  $e_2$  whose endpoints are  $u_1 \neq u_2$ . Let  $V_1, V_2, V_3$  be the vertex sets of the connected components of  $T \setminus \{e_1, e_2\}$  containing respectively  $u_1, v_1, u_2$  and let  $X_i = \phi^{-1}(V_i)$ . ■

## 3 Proof of the Splits-Equivalence Theorem

The heart of the nontrivial direction is an efficient algorithm for reconstructing phylogenies from splits known as *Tree Popping*.

- **Input:** A collection  $\Sigma = \{\sigma_1, \dots, \sigma_k\}$  pairwise compatible *X*-splits.
- **Output:** An *X*-tree  $\mathcal{T}$  such that  $\Sigma(\mathcal{T}) = \Sigma$ .

- *Initialization:* Let  $\mathcal{T}_0 = (T_0, \phi_0)$  be made of a single vertex labelled by  $X$ .
- *For  $i = 1 \dots k$ :*
  - \* Set  $\sigma_i = R_i|G_i$ .
  - \* Colour red (respectively green) the vertices of  $\mathcal{T}_{i-1} = (T_{i-1}, \phi_{i-1})$  in  $\phi_{i-1}(R_i)$  (respectively  $\phi_{i-1}(G_i)$ ).
  - \* Find the unique vertex  $w$  in  $\mathcal{T}_{i-1}$  such that all connected components of  $T_{i-1} \setminus \{w\}$  are monochromatic (i.e., all coloured vertices have the same colour).
  - \* Construct  $\mathcal{T}_i = (T_i, \phi_i)$  such that  $\Sigma(\mathcal{T}_i) = \{\sigma_1, \dots, \sigma_i\}$  by replacing  $w$  in  $\mathcal{T}_{i-1}$  with a new edge  $e = \{w_R, w_G\}$  as follows: the red (respectively green) components of  $T_{i-1} \setminus \{w\}$  are incident with  $w_R$  (respectively  $w_G$ ); the red (respectively green) labels of  $w$  are assigned to  $w_R$  (respectively  $w_G$ ).

The fact that the Tree Popping algorithm works is part of the proof of the Splits-Equivalence Theorem.

**Proof:**(of Splits-Equivalence Theorem) Assume  $\Sigma = \{\sigma_1, \dots, \sigma_k\}$  is a collection of pairwise compatible splits. We prove that  $\Sigma = \Sigma(\mathcal{T})$  for some  $\mathcal{T}$  by induction on  $k$ . The result is trivial for  $k = 0$ . Now assume that  $k \geq 1$  and that the result holds for  $k - 1$ . Hence, there is a unique  $\mathcal{T}_{k-1} = (T_{k-1}, \phi_{k-1})$  such that  $\Sigma(\mathcal{T}_{k-1}) = \{\sigma_1, \dots, \sigma_{k-1}\}$ . We apply the next lemma with  $\mathcal{T} = \mathcal{T}_{k-1}$  and  $\sigma = \sigma_k$ .

**LEM 2.8 (Label Painting Lemma)** *Let  $\mathcal{T} = (T, \phi)$  be an  $X$ -tree with  $T = (V, E)$ . Let  $\sigma = R|G$  be an  $X$ -split such that  $\sigma \notin \Sigma(\mathcal{T})$  but  $\sigma$  is compatible with all splits in  $\Sigma(\mathcal{T})$ . Colour red (respectively green) the vertices of  $\mathcal{T}$  in  $\phi(R)$  (respectively  $\phi(G)$ ). Then, there is a unique vertex  $w \in V$  such that the connected components of  $T \setminus \{w\}$  are monochromatic.*

**Proof:** *Existence of  $w$ .* We proceed by giving an orientation to each edge  $e = \{u, v\} \in E$ . By compatibility, there is exactly one component of  $T \setminus \{e\}$  which is monochromatic. Orient  $e$  away from that component. The resulting directed tree must have a *sink*  $w$ , that is, a vertex with no outgoing edge (indeed, start at any vertex and follow the directed edges; this process must stop in a finite tree otherwise there would be a cycle).

*Uniqueness of  $w$ .* Assume by contradiction that there are two such vertices  $w \neq w'$ . Choose an edge  $e$  on the path between  $w$  and  $w'$ . W.l.o.g., assume the compo-

ment of  $T \setminus \{e\}$  containing  $w$  is *not* monochromatic (by the compatibility assumption). Considering the component of  $T \setminus \{w'\}$  containing  $w$  gives a contradiction.

■

We now return to the proof of Theorem 2.6. Applying the procedure in the Tree Popping algorithm gives an  $X$ -tree  $\mathcal{T}_k$  identical to  $\mathcal{T}_{k-1}$  except for a single extra edge inducing the split  $\sigma_k$ . The uniqueness of  $\mathcal{T}_k$  follows from the uniqueness of  $\mathcal{T}_{k-1}$  and the uniqueness of  $w$ . ■

## 4 Applications of the Splits-Equivalence Theorem

We give two further applications of Theorem 2.6.

### 4.1 Refinement

The characterization of  $X$ -trees in terms of  $X$ -splits provides a natural partial order on the set of  $X$ -trees.

**DEF 2.9 (Refinement)** Let  $\mathcal{P}(X)$  be the set of  $X$ -trees. For  $\mathcal{T}, \mathcal{T}' \in \mathcal{P}(X)$ , we write  $\mathcal{T} \leq \mathcal{T}'$  when  $\Sigma(\mathcal{T}) \subseteq \Sigma(\mathcal{T}')$  and we say that  $\mathcal{T}'$  refines  $\mathcal{T}$ .

It can be checked that  $(\mathcal{P}(X), \leq)$  is a partial order. (Recall that a partial order on a set  $S$  is a relation  $\leq$  such that for all  $x, y, z \in S$ :

1. (Reflexivity)  $x \leq x$ .
2. (Antisymmetry) If  $x \leq y$  and  $y \leq x$  then  $x = y$ .
3. (Transitivity) If  $x \leq y$  and  $y \leq z$  then  $x \leq z$ .)

### 4.2 Splits metric

The symmetric difference  $A \Delta B$  of two sets  $A, B$  is the set  $(A \setminus B) \cup (B \setminus A)$ .

**DEF 2.10 (Splits Metric)** For a pair  $\mathcal{T}, \mathcal{T}' \in \mathcal{P}(X)$ , we let

$$d(\mathcal{T}, \mathcal{T}') = |\Sigma(\mathcal{T}) \Delta \Sigma(\mathcal{T}')|$$

be the splits metric between  $\mathcal{T}$  and  $\mathcal{T}'$ .

It can be checked that  $d$  is indeed a metric. Recall that a nonnegative function  $d$  on  $S \times S$  is a metric if for all  $x, y, z \in S$ :

1. (Definiteness)  $d(x, y) = 0$  if and only if  $x = y$ .
2. (Symmetry)  $d(x, y) = d(y, x)$ .
3. (Triangle Inequality)  $d(x, z) \leq d(x, y) + d(y, z)$ .

The first property follows from the Splits-Equivalence Theorem. The second property is immediate. For the triangle inequality

$$d(\mathcal{T}, \mathcal{T}'') \leq d(\mathcal{T}, \mathcal{T}') + d(\mathcal{T}', \mathcal{T}''),$$

note that if  $\sigma \in \Sigma(\mathcal{T})$  and  $\sigma \notin \Sigma(\mathcal{T}'')$ , then there are two cases: 1)  $\sigma \in \Sigma(\mathcal{T}')$  in which case  $\sigma$  contributes to  $d(\mathcal{T}', \mathcal{T}'')$ ; 2)  $\sigma \notin \Sigma(\mathcal{T}')$  in which case  $\sigma$  contributes to  $d(\mathcal{T}, \mathcal{T}')$ .

### 4.3 Consensus

For a collection  $\mathcal{P}$  of  $X$ -trees, let

$$\Sigma(\mathcal{P}) = \cup_{\mathcal{T} \in \mathcal{P}} \Sigma(\mathcal{T}),$$

and for  $\sigma \in \Sigma(\mathcal{P})$  let  $n_{\mathcal{P}}(\sigma)$  be the number of  $X$ -trees in  $\mathcal{P}$  that induce  $\sigma$ . Define

$$\Sigma_{1/2}(\mathcal{P}) = \left\{ \sigma \in \Sigma(\mathcal{P}) : \frac{n_{\mathcal{P}}(\sigma)}{|\mathcal{P}|} > \frac{1}{2} \right\}.$$

**THM 2.11 (Majority-Rule Consensus)**  $\Sigma_{1/2}(\mathcal{P})$  is the set of  $X$ -splits induced by a unique  $X$ -tree which is called the majority-rule consensus tree.

**Proof:** Take in two  $\sigma_1 \neq \sigma_2 \in \Sigma_{1/2}(\mathcal{P})$ . By construction, there must be  $\mathcal{T} \in \mathcal{P}$  such that  $\sigma_1$  and  $\sigma_2$  are both induced by  $\mathcal{T}$ . Hence they are pairwise compatible and we are done by the Splits-Equivalence Theorem. ■

### Further reading

The definitions and results discussed here were taken from Chapters 3 and 4 of [SS03]. Much more on the subject can be found in that excellent monograph. See also [SS03] for the relevant bibliographic references.

## References

- [SS03] Charles Semple and Mike Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.