Operads and phylogenetic trees

John Baez and Nina Otter

Berkeley, 4 November 2015

Bacteria
- Spirochetes
- Proteobacteria
- Cyanobacteria
- Planctomyces
- Bacteroides
- Cytophaga
- Thermotoga
- Aquifex
- Gram positives
  - Green Filamentous bacteria

Archaea
- Methanosarcina
- Methanobacterium
- Methanococcus
- T. celer
- Pyrodictium
- Halophiles
- Entamoebae

Eukaryota
- Slime molds
- Animals
- Fungi
- Plants
- Ciliates
- Flagellates
- Trichomonads
- Microsporidia
- Diplomonads

Berkeley, 4 November 2015
Trees are used in biology to give a graphic representation of hierarchical relationships such as:

- Evolutionary relationships among species or genes
- Migration patterns of species
- Histories of populations
- Spread of viral infections
- Evolution of language
Trees that represent the evolution of species or genes are called **phylogenetic trees**.

In the past, phylogenetic trees were constructed using morphological similarities among the species.
Trees that represent the evolution of species or genes are called phylogenetic trees.

In the past, phylogenetic trees were constructed using morphological similarities among the species.

Nowadays biologists use molecular sequences such as DNA or RNA.
Such trees are usually rooted and semi-labelled:

The construction of phylogenetic trees can be subdivided into roughly four steps:

- Alignment of sequences
- Selection of a mathematical model to describe sequence evolution
- Selection of a tree-building method
- Assessment of the quality of the resulting tree
Models of DNA evolution are usually Markov chain models of nucleotide substitution.

The simplest of these models is the Jukes-Cantor model: it assumes that at any time the nucleotides have the same frequency and that they have an equal substitution rate.

The thickness of the arrows represents the substitution rate of the nucleotides.

The size of the circles represents the frequency at equilibrium.
A simple model for tree reconstruction is the least-square model:

- First compute the evolutionary distances $d_{ij}$ between any two sequences $i$ and $j$

- Find the tree that minimizes the following expression:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} (\hat{d}_{ij} - d_{ij})^2$$

where $\hat{d}_{ij}$ is the branch length between the nodes representing sequence $i$ and sequence $j$. 
The resulting tree is often just a good approximation: uncertainty about branching order and relationship among leaves is sometimes represented by filling out the uncertain regions of the tree, as e.g.
In 2001 Billera, Holmes and Vogtmann put forward a geometric model to parametrize the space of trees with fixed set of leaf labels.

They define a **phylogenetic $n$-tree** to be a rooted tree with leaves labelled by natural numbers $1, \ldots, n$ and internal edges labelled by numbers in $(0, \infty)$, e.g.:

![Diagram of a phylogenetic $n$-tree with labels 0, 1, 2, 3, 4, 5, and $t_3 > 0$.]

- The root is labelled 0, the leaves 1, 2, 3, 4, 5.
- Where $t_3 > 0$. 

Nina Otter (Oxford)  Phylogenetic operad  Berkeley, 4 November 2015  9 / 40
The embedding of the tree in the plane is irrelevant, so these are the same phylogenetic tree:
A tree $T$ with $r$ internal edges is represented by a point in the $r$-dimensional orthant $(0, \infty)^r$. 

\[(1, 1) = \bullet\]
A tree $T$ with $r$ internal edges is represented by a point in the $r$-dimensional orthant $(0, \infty)^r$.

To each other point in the orthant associate the tree that is combinatorially the same as $T$ but with different edge lengths.
A tree $T$ with $r$ internal edges is represented by a point in the $r$-dimensional orthant $(0, \infty)^r$.

To each other point in the orthant associate the tree that is combinatorially the same as $T$ but with different edge lengths.

To points on the boundary of the orthant assign trees obtained from $T$ by shrinking some interior edge of $T$ to 0.
The space $\mathcal{T}_n$ of phylogenetic $n$-trees is given by taking an $(n-2)$-orthant for each binary $n$-tree and gluing them along their common faces. For example this is $\mathcal{T}_3$:
BHV put a metric on the space of phylogenetic trees by setting the distance between two points in the same orthant to be the Euclidean distance, and between points in different orthants to be the length of the shortest piecewise path between the points.

It was recently suggested that this space is *the* space for statistical analysis of phylogenetic trees. ¹

---

¹ *Point estimates in phylogenetic reconstructions, P. Benner, M. Bačák, P-Y. Bourguignon*, 2014
BHV put a metric on the space of phylogenetic trees by setting the distance between two points in the same orthant to be the Euclidean distance, and between points in different orthants to be the length of the shortest piecewise path between the points.

It was recently suggested that this space is the space for statistical analysis of phylogenetic trees. ¹

On the other hand, this space arises naturally in operad theory!

¹ Point estimates in phylogenetic reconstructions, P. Benner, M. Bačák, P-Y. Bourguignon, 2014
The prototypical example of operad is the **endomorphism operad**: Let $X$ be a set, and let $\text{End}_X(n) = \text{Maps}(X^n, X)$.
The prototypical example of operad is the **endomorphism operad**: Let $X$ be a set, and let $\text{End}_X(n) = \text{Maps}(X^n, X)$.

We can think of $f \in \text{End}_X(n)$ as a box with $n$ incoming wires and one outgoing wire. For example this is $f \in \text{End}_X(3)$:
We can compose such maps together, e.g. like this:

We denote such a composite by $g \circ_2 f$. It is an element of $\text{End}_X(\mathcal{X})$. 
The order of composition doesn’t really matter, so this sort of composition is unambiguous:
The order of composition doesn’t really matter, so this sort of composition is unambiguous:

Furthermore, $id_X : X \to X$ is an identity for this composition.
An **operad** $O$ (in sets) is a collection of sets $O(0), O(1), \ldots$ together with composition maps

$$\circ_i : O(n) \times O(m) \to O(n + m - 1) \quad i = 1, \ldots, n$$

and a unit element

$$1_O \in O(1)$$

such that the composition maps are associative and unital.

An operad is **symmetric** if there is an action $S_n \times O(n) \to O(n)$ and the composition maps are equivariant.

One can define operads in any symmetric monoidal category, such as $(\text{Top}, \times)$ or $(\text{Vect}_k, \otimes)$. We call this topological operads, linear operads, etc.

All operads here will be topological and symmetric.
The *reason d'être* of operads is that they have algebras:

An **algebra** over an operad $O$ is a topological space $A$ together with a continuous map

$$f : A^n \rightarrow A$$

for every $f \in O(n)$. These maps satisfy some associativity and unitality conditions.

A **coalgebra** over $O$ is a topological space $A$ together with continuous maps

$$f : A \rightarrow A^n$$

satisfying coassociativity and counitality conditions.
Some operads and their (co)algebras:

- $\text{Com}(n) = \star$ for all $n \geq 0$, its algebras are the commutative topological monoids, its coalgebras are the cocommutative topological coalgebras with counit.

- $\text{Com}(n) = \star$ for all $n \geq 1$, its algebras are the commutative topological semigroups, its coalgebras are the cocommutative topological coalgebras, not necessarily with counit.
Define a **phylogenetic $n$-tree** to be a rooted tree with leaves labelled by numbers $1, 2, \ldots, n$, root labelled by $0$ and edges labelled by numbers in $[0, \infty)$. We require that:

- the length of every edge is positive, except perhaps for edges incident to a leaf or the root;
- no 2-ary vertices are allowed.
For example, here is a phylogenetic tree with 5 leaves:

```
  3  1  4  5  2
   \  /    /
    \  /  t5
     \ /  \
      \ /
       \t4
        /
       t3
      /  \
    t7 /    \
      /      \
     /        \
    /          \
   /            \
  0--------------
```

where $t_3 > 0$ and $t_1, t_2, t_4, t_5, t_6, t_7 \geq 0$. The root is labelled 0, the leaves 1, 2, 3, 4, 5.

The embedding of a phylogenetic tree in the plane is irrelevant.
Note that:

- the trees in BHV have all edges with two adjacent vertices, but it doesn’t make a difference to consider external edges with only one adjacent vertex, like this:

```
  1  3  4  5  2
   \   \   \   \
    \   \   \
     \   \\  
      \  \\
       \ 
        0
```

- They do not label the external edges; doing that would consist in taking $\mathcal{T}_n \times [0, \infty)^{n+1}$, so all the interesting topology lies in $\mathcal{T}_n$. 
Denote by $\text{Phyl}_n$ the set of phylogenetic $n$-trees.

**Proposition.** (JB, NO 2015) There is an operad $\text{Phyl}$, the **phylogenetic operad**, whose set of $n$-ary operations is $\text{Phyl}_n$.

What is this operad?

The coproduct of two operads $O$ and $O'$ is an operad whose $n$-ary operations are equivalence classes of rooted trees with $n$-leaves and $k$-ary vertices labelled by $k$-ary operations in $O$ and $O'$.

$\text{Phyl}$ is the coproduct of $\text{Com}$ and $[0,\infty)$, where:

- $\text{Com}$ is the operad with $\text{Com}_n$ given by the one-point set for $n > 0$ and the empty set for $n = 0$,

- $[0,\infty)$ is the operad having only unary operations, one for each $t \in [0,\infty)$, with composition of operations given by addition.
If we endow the BHV space of trees with the topology induced by their metric and $\text{Phyl}_n$ with the topology of the coproduct, we have:

**Proposition.** (JB NO 2015) For all $n > 1$ there is a homeomorphism

$$\text{Phyl}_n \cong \mathcal{T}_n \times [0, \infty)^{n+1}$$
If we endow the BHV space of trees with the topology induced by their metric and \( \text{Phyl}_n \) with the topology of the coproduct, we have:

**Proposition.** (JB NO 2015) For all \( n > 1 \) there is a homeomorphism

\[
\text{Phyl}_n \cong \mathcal{T}_n \times [0, \infty)^{n+1}
\]

But there is more to this story:
A **continuous time Markov process** on a finite set $X$ is a collection of linear maps

$$\alpha(t): \mathbb{R}^X \rightarrow \mathbb{R}^X \quad t \in [0, \infty)$$

such that:

1. $\alpha(t + s) = \alpha(s) \circ \alpha(t)$ for all $t, s \in [0, \infty)$
2. $\alpha(0) = \text{id}_{\mathbb{R}^X}$
3. $\alpha$ depends continuously on $t$
4. if $v \in \mathbb{R}^X$ is a probability distribution on $X$, so is $\alpha(t)(v)$ for all $t$. 
A **continuous time Markov process** on a finite set $X$ is a collection of linear maps 

$$\alpha(t): \mathbb{R}^X \to \mathbb{R}^X \quad t \in [0, \infty)$$

such that:

1. $\alpha(t + s) = \alpha(s) \circ \alpha(t)$ for all $t, s \in [0, \infty)$
2. $\alpha(0) = id_{\mathbb{R}^X}$
3. $\alpha$ depends continuously on $t$
4. if $v \in \mathbb{R}^X$ is a probability distribution on $X$, so is $\alpha(t)(v)$ for all $t$.

In other words, $\mathbb{R}^X$ is an algebra over $[0, \infty)$ in $\text{Top}$ satisfying condition (4).
A **continuous time Markov process** on a finite set $X$ is a collection of linear maps

$$\alpha(t): \mathbb{R}^X \to \mathbb{R}^X \quad t \in [0, \infty)$$

such that:

1. $\alpha(t + s) = \alpha(s) \circ \alpha(t)$ for all $t, s \in [0, \infty)$
2. $\alpha(0) = id_{\mathbb{R}^X}$
3. $\alpha$ depends continuously on $t$
4. If $v \in \mathbb{R}^X$ is a probability distribution on $X$, so is $\alpha(t)(v)$ for all $t$.

In other words, $\mathbb{R}^X$ is an algebra over $[0, \infty)$ in $\text{Top}$ satisfying condition (4).

Since addition in $[0, \infty)$ is commutative, this is the same as a coalgebra over $[0, \infty)$ satisfying condition (4).
$\mathbb{R}^X$ is naturally a coalgebra over $\text{Com}$: the unique $n$-ary operation $\mu_n$ acts by sending $\nu \in \mathbb{R}^X$ to $(\nu, \ldots, \nu)$ in $\mathbb{R}^X \otimes \cdots \otimes \mathbb{R}^X$. 
\( \mathbb{R}^X \) is naturally a coalgebra over \( \text{Com} \): the unique \( n \)-ary operation \( \mu_n \) acts by sending \( \nu \in \mathbb{R}^X \) to \( (\nu, \ldots, \nu) \) in \( \mathbb{R}^X \otimes \cdots \otimes \mathbb{R}^X \).

By general abstract non-sense a coalgebra over \( \text{Com} + [0, \infty) \) is a topological space \( A \) that is both a coalgebra of \( \text{Com} \) and \( [0, \infty) \).
\( \mathbb{R}^X \) is naturally a coalgebra over \( \text{Com} \): the unique \( n \)-ary operation \( \mu_n \) acts by sending \( \nu \in \mathbb{R}^X \) to \( (\nu, \ldots, \nu) \) in \( \mathbb{R}^X \otimes \cdots \otimes \mathbb{R}^X \).

By general abstract non-sense a coalgebra over \( \text{Com} + [0, \infty) \) is a topological space \( A \) that is both a coalgebra of \( \text{Com} \) and \( [0, \infty) \).

We thus obtain:

**Proposition.** (JB, NO 2015) Given a Markov process on a finite set \( X \), the vector space \( \mathbb{R}^X \) naturally becomes a coalgebra of \( \text{Phyl} \).
Furthermore, Phyl is closely related to an operad used in algebraic topology. This is the $W$-construction of Boardman and Vogt applied to the operad $\text{Com}$. 

For any operad $O$ operations in $W(O)_n$ are equivalence classes of rooted planar trees with internal edges labelled by numbers in $(0, 1]$ and external edges labelled by 1, vertices labelled by operations in $O$. The equivalence relation comes from permuting edges coming into a vertex:
Furthermore, Phyl is closely related to an operad used in algebraic topology. This is the $W$-construction of Boardman and Vogt applied to the operad $Com$.

For any operad $O$ operations in $W(O)_n$ are equivalence classes of rooted planar trees with internal edges labelled by numbers in $(0, 1]$ and external edges labelled by 1, vertices labelled by operations in $O$. The equivalence relation comes from permuting edges coming into a vertex:

$$T_1 \ldots T_r \sim T_{\sigma^{-1}(1)} \ldots T_{\sigma^{-1}(r)}$$

for $\sigma \in S_r$. 

$$\begin{align*}
T_1 & \quad \cdots \quad T_r \\
\downarrow & \quad \downarrow & \quad \downarrow \\
\bullet & \quad \sigma & \quad f
\end{align*}$$

$$\begin{align*}
T_{\sigma^{-1}(1)} & \quad \cdots \quad T_{\sigma^{-1}(r)} \\
\downarrow & \quad \downarrow \\
\bullet & \quad f
\end{align*}$$
Here is an operation in $W(O)_4$:

where $t_1, t_2, t_3 > 0$ and $i \in O(0), h \in O(3), g, f \in O(2)$. 
Composition is given by grafting trees and assigning length 1 to the new edges arising.
Composition is given by grafting trees and assigning length 1 to the new edges arising.

$W(O)_n$ has an obvious topology, induced by the topology of the spaces in $O$ and the topology of $[0, 1]$. 
Berger and Moerdijk showed that if $O$ is well-pointed and the action of the symmetric groups is free, then $W(O)$ is a cofibrant replacement of $O$ in the model category on the category of topological operads induced by the Serre monoidal model structure on spaces.

$W(Com)$ is not cofibrant, because the action of the symmetric group $S_n$ on $W(Com)_n$ is trivial.

Algebras over $W(Com)$ are commutative $A_\infty$ algebras: associativity holds up to coherent homotopy, but commutativity is strict.

However, $W(Com)$ is still interesting:
Berger and Moerdijk showed that if $O$ is well-pointed and the action of the symmetric groups is free, then $W(O)$ is a cofibrant replacement of $O$ in the model category on the category of topological operads induced by the Serre monoidal model structure on spaces.

$W(\text{Com})$ is not cofibrant, because the action of the symmetric group $S_n$ on $W(\text{Com})_n$ is trivial.
Berger and Moerdijk showed that if $O$ is well-pointed and the action of the symmetric groups is free, then $W(O)$ is a cofibrant replacement of $O$ in the model category on the category of topological operads induced by the Serre monoidal model structure on spaces.

$W(\text{Com})$ is not cofibrant, because the action of the symmetric group $S_n$ on $W(\text{Com})_n$ is trivial.

Algebras over $W(\text{Com})$ are commutative $A_\infty$ algebras: associativity holds up to coherent homotopy, but commutativity is strict.
Berger and Moerdijk showed that if $O$ is well-pointed and the action of the symmetric groups is free, then $W(O)$ is a cofibrant replacement of $O$ in the model category on the category of topological operads induced by the Serre monoidal model structure on spaces.

$W(\text{Com})$ is not cofibrant, because the action of the symmetric group $S_n$ on $W(\text{Com})_n$ is trivial.

Algebras over $W(\text{Com})$ are commutative $A_\infty$ algebras: associativity holds up to coherent homotopy, but commutativity is strict.

However, $W(\text{Com})$ is still interesting:
Consider the closed interval $[0, \infty]$ and define addition on $[0, \infty]$ in the obvious way, where
\[ \infty + t = t + \infty = \infty. \]
Then $[0, \infty]$ becomes a topological monoid, so there is an operad with only unary operations, one for each $t \in [0, \infty]$. Let’s call this operad $[0, \infty]$. 

Instead of labelling edges of trees by numbers in $[0, 1]$ we can adapt the Boardman Vogt resolution by labelling edges by numbers in $[0, \infty]$. 

We have:

**Proposition.** (JB, NO 2015) For any operad $O$ the operad $W(O)$ is a non-unital suboperad of $O + [0, \infty]$ and the spaces $W(O)^n$ and $O^n + [0, \infty]^n$ are homotopy equivalent, if $n \neq 1$.

Moreover, for every $n > 1$ there is a homeomorphism $W(Com)^n \sim = T^n$. 

Nina Otter (Oxford) Phylogenetic operad Berkeley, 4 November 2015 33 / 40
Consider the closed interval $[0, \infty]$ and define addition on $[0, \infty]$ in the obvious way, where
\[
\infty + t = t + \infty = \infty.
\]
Then $[0, \infty]$ becomes a topological monoid, so there is an operad with only unary operations, one for each $t \in [0, \infty]$. Let’s call this operad $[0, \infty]$.

Instead of labelling edges of trees by numbers in $[0, 1]$ we can adapt the Boardman Vogt resolution by labelling edges by numbers in $[0, \infty]$. 
Consider the closed interval \([0, \infty]\) and define addition on \([0, \infty]\) in the obvious way, where

\[
\infty + t = t + \infty = \infty.
\]

Then \([0, \infty]\) becomes a topological monoid, so there is an operad with only unary operations, one for each \(t \in [0, \infty]\). Let’s call this operad \([0, \infty]\).

Instead of labelling edges of trees by numbers in \([0, 1]\) we can adapt the Boardman Vogt resolution by labelling edges by numbers in \([0, \infty]\).

We have:

**Proposition.** (JB, NO 2015) For any operad \(O\) the operad \(W(O)\) is a non-unital suboperad of \(O + [0, \infty]\) and the spaces \(W(O)_n\) and \(O + [0, \infty]_n\) are homotopy equivalent, if \(n \neq 1\).
Consider the closed interval \([0, \infty]\) and define addition on \([0, \infty]\) in the obvious way, where
\[
\infty + t = t + \infty = \infty.
\]
Then \([0, \infty]\) becomes a topological monoid, so there is an operad with only unary operations, one for each \(t \in [0, \infty]\). Let’s call this operad \([0, \infty]\).

Instead of labelling edges of trees by numbers in \([0, 1]\) we can adapt the Boardman Vogt resolution by labelling edges by numbers in \([0, \infty]\).

We have:

**Proposition.** (JB, NO 2015) For any operad \(O\) the operad \(W(O)\) is a non-unital suboperad of \(O + [0, \infty]\) and the spaces \(W(O)_n\) and \(O + [0, \infty]_n\) are homotopy equivalent, if \(n \neq 1\).

Moreover, for every \(n > 1\) there is a homeomorphism \(W(\text{Com})_n \cong \mathcal{T}_n\).
And the larger operad $\text{Com} + [0, \infty]$, a compactification of $\text{Phyl} = \text{Com} + [0, \infty)$, is also interesting.
And the larger operad $\text{Com} + [0, \infty]$, a compactification of $\text{Phyl} = \text{Com} + [0, \infty)$, is also interesting.

The reason is that any Markov process $\alpha : [0, \infty) \to \text{End}(\mathbb{R}^X)$ approaches a limit as $t \to \infty$. Indeed, it extends uniquely to a homomorphism of topological monoids $\alpha : [0, \infty] \to \text{End}(\mathbb{R}^X)$. 
And the larger operad $\text{Com} + [0, \infty]$, a compactification of $\text{Phyl} = \text{Com} + [0, \infty)$, is also interesting.

The reason is that any Markov process $\alpha: [0, \infty) \to \text{End}(\mathbb{R}^X)$ approaches a limit as $t \to \infty$. Indeed, it extends uniquely to a homomorphism of topological monoids $\alpha: [0, \infty] \to \text{End}(\mathbb{R}^X)$.

We thus get:

**Corollary.** (JB NO 2015) Given a Markov process on a finite set $X$, the vector space $\mathbb{R}^X$ naturally becomes a coalgebra of $\text{Com} + [0, \infty]$. 
Finally.. some mysteries!
1) The mystery of tropical trees!
The **tropical semiring** is $\mathbb{R} \cup \{-\infty\}$ with addition given by taking the maximum and multiplication given by the classical addition. We can do algebraic geometry over this semiring and define ‘tropical curves’.

In 2007, Gathmann, Kerber and Markwig showed that a certain moduli space of genus 0 tropical curves with $n + 1$ marked points is the space of trees $T_n$ studied by Billera, Holmes and Vogtmann.

Also in 2007, Mikhalkin showed this moduli space has a compactification that is a smooth compact tropical variety.

**Proposition.** (NO 2014) This compactification is $W(\text{Com})_n$. The operad structure on $W(\text{Com})$ corresponds to the ‘tropical clutching map’ described by Abramovich, Caporaso and Payne in 2012.

**The mystery:** Is this good for something? What’s $W(\text{Com})$ for curves of higher genus?
2) *The mystery of horizontal gene transfer!*

![Diagram showing the relationship between Bacteria, Archaea, and Eukaryotes, highlighting the common ancestral community of primitive cells with horizontal gene transfer events labeled 1 to 6.]
In 2013 Chan, Carlsson and Rabadan studied horizontal gene transfer using persistent homology.

They take a set of genomes and calculate the genetic distances between pairs of sequences. Using this distance as parameter they calculate persistent homology.

Zero dimensional homology provides information about vertical evolution. One dimensional homology provides information about horizontal evolution.

They conjecture that higher dimensional homology results from multiple horizontal exchanges. Representatives of higher dimensional classes are polytopes.

**The mystery: How can we parametrise these structures? Would a ‘higher dimensional analogue’ of the operad \( \text{Phyl} \) be any good?**
References

- *Operads and phylogenetic trees*, John Baez and Nina Otter, 2015 (soon on the arXiv!)
- *The Boardman-Vogt resolution of operads in monoidal model categories*, Clemens Berger and Ieke Moerdijk, 2005