Using Category Theory to justify EvoHop

David Aaby

July 26, 2007

Abstract

This paper seeks to justify a process that finds importance ranking for bacteriophage proteins and creates evolutionary distances between phage in order to construct a phylogenetic tree. Category theory is used to define two categories; one category to represent distance graphs and one category to represent reversible Markov chains. We also find functors to map between categories in order to show that EvoHop is a natural process from transforming evolutionary protein distances to evolutionary phage distances.

1 Introduction

Given a population of organisms, a quantity of interest to phylogenetics is the pairwise evolutionary distance between organisms. There are many ways to calculate the evolutionary distance between organisms. The simplest way is to look at one protein that is common in all of the organisms, sequence the DNA code for this protein, and count the number of mutations or dissimilarites between this genetic code in each organism. The more dissimilar the DNA code is for each organism, the greater the evolutionary distance between them. This evolutionary protein distance is used to create evolutionary distances between organisms [6]. Generally, protein distances for phage are obtained using a database such as ProtDist. In our research, we are looking specifically at mutations between bacteriophage in hopes of creating a phylogenetic tree.

EvoHop was the name given to the random walk on a set of phage proteins. EvoHop accomplishes two tasks. First, it determines an importance ranking for all phage proteins within the set. This is done using a method similar to Google's PageRank algorithm. It is important to rank all proteins within all phages of the community because there is no single protein that all phages have in common. This is why alternative phylogenetic techniques must be used in order to create a tree for phage.

Second, EvoHop calculates evolutionary distances between phages from evolutionary distances between proteins. We can then use these distances in order to create a phylogenetic tree for phage, which will show how the phages are related to each other through evolution. Starting with protein distance data, an algorithm was used to transform the distances into transition probabilities between proteins. The transition probabilities represent the probability that one protein will mutate into another. Then, Markov chain lumping techniques were used in order to group together all of the proteins that were contained in each phage. The lumping algorithm transformed the protein-protein transition probabilities into phage-phage transition probabilites. Within reversible Markov chains, there exists a well-established, natural way to lump [3]. Then, another algorithm was used to create distances between phages, which can then be used to construct a phylogenetic tree [1].

Using category theory, we hope to be able to justify that this process is a natural or canonical way of creating intra-phage distances for the use in phylogenetics. We aim to justify a method of of transforming symmetric matrices, which represent evolutionary protein distances or phage distances, to stochastic matrices for a reversible Markov chain, which represent transition probabilities between proteins or transition probabilities between phage, in order to model the phage evolution. We will use *Category Theory* as it provides unity in abstract mathematical structures and a well-defined notion of *natural* transformations.

2 Basic Terms and Definitions

Definition 2.1. *Let S be a set such that $|S| = n$ for some integer n. Let P be a partition of S such that $P = \{\phi_1, \phi_2, ..., \phi_k\}$. A* refinement of a partition P*, denoted $\hat{P}$, is defined such that $\hat{P} \supset P$ if for each $\phi_i \in P \; \exists \; \phi_{i1}, \phi_{i2}, ..., \phi_{iz}$ such that $\{\phi_{i1}, \phi_{i2}, ..., \phi_{iz}\}$ is a partition of ϕ_i.*

Example 2.2. *Let $S = \{1, 2, 3, 4, 5, 6, 7, 8\}$*
$P = \{\{1, 2\}, \{3, 4, 5, 6, 7, 8\}\}$
$\hat{P} = \{\{1, 2\}, \{3, 4\}, \{5, 6, 7, 8\}\}$
$\hat{P}$ is a refinement *of P.*

Definition 2.3. *A* commutative diagram *is a diagram of objects and morphisms such that, when selecting two objects, one can follow any path through the diagram and achieve the same result by composition [7].*

Example 2.4. *Let A, B, C, D be sets and let w, x, y, z be functions.*

$$\begin{tikzcd}
A \arrow[r, "w"] \arrow[d, "z"'] & C \arrow[d, "x"] \\
B \arrow[r, "y"'] & D
\end{tikzcd}$$

The diagram is commutative if $y \circ w = z \circ x$.

Definition 2.5. *A Category is an ordered triple $(C, hom(C), \circ)$ where:*

- *C is a set of objects,*

- *hom(C) is a set of* morphisms *where each morphism assigns a unique (source) object in C to a (target) object in C,*

- *and $\circ$ is a binary operation called* composition of morphisms.

A category is also governed by two axioms:

- *associativity*

  If $f : a \rightarrow b, g : b \rightarrow c$, and $h : c \rightarrow d$ are morphisms then:

  $$(f \circ g) \circ h = f \circ (g \circ h) \tag{1}$$

- *identity*

  For each object x there is a unique identity $1_x : x \rightarrow x$ such that if f is a morphism from a to b then:

  $$1_b \circ f = f = f \circ 1_a \tag{2}$$

[5]

Within the category, if the objects are sets and the morphisms are functions, then identity and associativity arise automatically, and need not be formally proven.

Definition 2.6. *A* natural transformation *is any mapping between categories that can be proven using commutative diagrams. [7]*

Our two categories will be small categories that are designed for a very specific purpose: to justify that EvoHop is a natural process for creating evolutionary distances between phage.

3 Category *Sym*

Our first category, denoted *Sym*, will be defined on the set of symmetric matrices with non-negative real entries and corresponding column partitions. Each symmetric matrix represents a distance graph or edge-weighted graph such that each weight is non-negative and satisfies the triangle inequality. Each partition represents the proteins found in one particular phage.

We define the morphism $f : (D_1, P_1) \rightarrow (D_2, P_2)$ as follows:

Definition 3.1. *Let S be a set such that $|S| = n$. Let D be a $n \times n$ symmetric matrix with non-negative real entries, and let P_1 be the trivial partition of S such that each subset of P_1 contains only one element. Let $P_2 = \{\phi_1, \phi_2, ..., \phi_k\}$ be a partition of P_1 . The morphism f maps (D_1, P_1) to (D_2, P_2) where D_2 is the $k \times k$ matrix such that*

$$(D_2)_{ij} = \frac{|\phi_i| \times |\phi_j|}{\sum\limits_{\ell \in \phi_i, m \in \phi_j} \frac{1}{D_{\ell m}}} \tag{3}$$

Essentially, the morphism f determines the harmonic mean distance between subsets of the partition P.

Theorem 3.2. Sym *is a category.*

Proof. In order for *Sym* to be a category, we need to show that the identity and associativity of morphisms hold, and also that the morphisms within the category commute.

identity: Let $f : (D_1, P_1) \to (D_2, P_2)$ be our morphism as previously defined. This morphism always creates a trivial partition, where each subset of the partition contains only one element.

Let $1_{(x,y)} : (x, y) \to (x, y)$ be an identity morphism.

It must hold that $1_{(D_2,P_2)} \circ f = f = f \circ 1_{(D_1,P_1)}$ where $1_{(D_1,P_1)}$ creates a trivial partition of D such that each subset contains only one element (i.e. one protein) of D. Similarly, $1_{(D_2,P_2)}$ creates a trivial partition of D_1 such that each subset contains only one element (i.e. one phage) of D_2.

Thus, ∃ an identity morphism ∀ $(x, y) \in (D, P)$

associativity: Let f, g, and h be morphisms where f is defined as above.

We need to show that $(h \circ g) \circ f = h \circ (g \circ f)$. Since P groups together all proteins within the same phage, no further partitioning is possible. Since f creates a trivial partition where each partition contains only one element, no further partitioning is possible. This implies that g and h must be identity morphisms.

So $(h \circ g) \circ f = g \circ f = f = g \circ f = h \circ (g \circ f)$.

Therefore, associativity holds in our class of morphisms.

commutativity: Let $P_1 = \{\{1\}, \{2\}, ...\{n\}\}$. Let D_1 be a $n \times n$ matrix.

Let P_2 be a partition of P_1 where $P_2 = \{\phi_1, \phi_2, ..., \phi_k\}$.

Let D_2 be a $k \times k$ matrix such that

$$(D_2)_{ij} = \frac{|\phi_i| \times |\phi_j|}{\sum\limits_{\ell \in \phi_i, m \in \phi_j} \frac{1}{(D_2)_{\ell m}}}$$

Let P_3 be a refinement of P_2 where $P_3 = \{\theta_1, \theta_2, ..., \theta_p\}$ such that $p \leq k$.

Then D_3 is a $p \times p$ matrix such that

$$(D_3)_{ij} = \frac{|\theta_i| \times |\theta_j|}{\sum\limits_{\ell \in \theta_i, m \in \theta_j} \frac{1}{(D_3)_{\ell m}}}$$

Let f be a morphism such that $f : (D_1, P_1) \to (D_2, P_2)$. Let g be a morphism such that $g : (D_2, P_2) \to (D_3, P_3)$.

Since the elements of P_1 are contained in each ϕ_i and each element ϕ_i of P_2 is contained in each element θ_j for some $j \in P_3$, this implies that the elements of P_1 are contained in the elements of P_3. It must follow that ∃ a morphism h such that

$h : (D_1, P_1) \to (D_3, P_3)$.

Therefore, the morphism is commutative.

Therefore, *Sym* is a category.

□

The proof can be shown by the following commutative diagram:

$$\begin{tikzcd}
(D_1, P_1) \arrow[d, "h"'] \arrow[dr, "f"] & \\
(D_3, P_3) & (D_2, P_2) \arrow[l, "g"]
\end{tikzcd}$$

Figure 1: A commutative diagram for the category *Sym*.

Example 3.3. *Given a symmetric distance matrix D and partition P, find L_D.*

$$D = \begin{bmatrix} - & 1.4 & \infty & .78 & \infty & \infty & .85 \\ 1.4 & - & \infty & 2.3 & \infty & \infty & 1.1 \\ \infty & \infty & - & \infty & 1.6 & 2.9 & \infty \\ .78 & 2.3 & \infty & - & \infty & \infty & .37 \\ \infty & \infty & 1.6 & \infty & - & .43 & \infty \\ \infty & \infty & 2.9 & \infty & .43 & - & \infty \\ .85 & 1.1 & \infty & .37 & \infty & \infty & - \end{bmatrix}$$

$S = \{1, 2, 3, 4, 5, 6, 7\}$
$P = \{\{1, 5\}, \{2, 3, 4\}, \{6, 7\}\}$

$$(L_D)_{1,2} = (L_D)_{\{1,5\},\{2,3,4\}} = \frac{2\times3}{\frac{1}{D_{1,2}}+\frac{1}{D_{1,3}}+\frac{1}{D_{1,4}}+\frac{1}{D_{5,2}}+\frac{1}{D_{5,3}}+\frac{1}{D_{5,4}}} = \frac{6}{\frac{1}{1.4}+\frac{1}{\infty}+\frac{1}{.78}+\frac{1}{\infty}+\frac{1}{1.6}+\frac{1}{\infty}} = 2.289$$

$$(L_D)_{1,3} = (L_D)_{\{1,5\},\{6,7\}} = \frac{2\times2}{\frac{1}{D_{1,6}}+\frac{1}{D_{1,7}}+\frac{1}{D_{5,6}}+\frac{1}{D_{5,7}}} = \frac{4}{\frac{1}{\infty}+\frac{1}{.85}+\frac{1}{.43}+\frac{1}{\infty}} = 1.142$$

$$L_D = \begin{bmatrix} - & 2.289 & 1.142 \\ 2.289 & - & 1.516 \\ 1.142 & 1.516 & - \end{bmatrix}$$

4 Category *Sto*

Our second category, denoted *Sto*, will be defined on the set of stochastic matrices whose columns sum to 1 and the corresponding column partitions. Each stochastic matrix represents a reversible Markov chain and each partition represents the proteins found in a particular phage. We define our morphism $L : (M_1, P_1) \to (M_2, P_2)$ as follows:

Definition 4.1. *Let S be a set such that $|S| = n$. Let P_1 be a trivial partition of S such that each subset of P_1 contains only one element. Let M_1 be a column stochastic matrix and let $P_2 = \{\phi_1, \phi_2, ..., \phi_n\}$ be a partition of P_1. The morphism L maps (M_1, P_1) to (M_2, P_2) where M_2 is the lumped matrix such that $M_2 = CM_1D$ where C and D are the collect and distribute matrices, respectively.*

Definition 4.2. *Let C be the $m \times n$* collect *matrix where the i^{th} row is a vector with $1's$ in the components corresponding to states in M_i and $0's$ otherwise [3].*

Definition 4.3. *Let D be the $n \times m$* distribute *matrix whose j^{th} column is the probability vector having equal components for states in M_j and 0 otherwise [3].*

Theorem 4.4. Sto *is a category.*

Proof. Again, we need to show that the identity and associativity of morphisms hold, and also that the morphisms commute within the category.

identity: Let $L : (M_1, P_1) \to (M_2, P_2)$ be our morphism as defined above. The morphism L always creates a trivial partition, where each partition contains only one element.

Let $1_{(x,y)} : (x, y) \to (x, y)$ be an identity morphism.

It must hold that $1_{(M_2,P_2)} \circ L = L = L \circ 1_{(M_1,P_1)}$ where $1_{(M_1,P_1)}$ creates a trivial partition where each element (i.e. each protein) is lumped together with only itself. Similarly, $1_{(M_2,P_2)}$ creates a trivial partition where each element (i.e. each phage) is lumped together with only itself.

Thus $\exists$ an identity morphism $\forall$ $(x, y) \in (M, P)$

associativity: Let L, Q, and R be morphisms where L is defined as above.

We need to show that $(R \circ Q) \circ L = R \circ (Q \circ L)$. Since P lumps together all proteins within the same phage, no further partitioning is possible. Since L creates a trivial partition where each partition contains only one element, no further partitioning (i.e. lumping) is possible. This implies that Q and R must be identity morphisms.

So $(R \circ Q) \circ L = R \circ L = L = Q \circ L = R \circ (Q \circ L)$.

Therefore, associativity holds in our class of morphisms.

commutativity: Let $P_1 = \{1, 2, ..., n\}$

Let M_1 be a $n \times n$ column stochastic matrix.

Let P_2 be a partition of P_1 such that $P_2 = \{\phi_1, \phi_2, ..., \phi_k\}$

Let M_2 be a $k \times k$ column stochastic matrix such that

$$M_2 = C_1 M_1 D_1$$

where C_1 and D_1 are the collect and distribute matrices of M_1, respectively.

Let P_3 be a refinement of P_2 such that $P_3 = \{\theta_1, \theta_2, ..., \theta_p\}$ where $p \leq k$.

Let M_3 be a $p \times p$ matrix such that

$$M_3 = C_2 M_2 D_2$$

Let f and g be morphisms within *Sto* such that $f : (M_1, P_1) \to (M_2, P_2)$ and $g : (M_2, P_2) \to (M_3, P_3)$.

Since the elements of P_1 are contained within each ϕ_i of P_2 and each element ϕ_i is contained in each element θ_j for some $j \in P_3$, this implies that the elements of P_1 are contained in the elements of P_3.

Therefore, there must exist a morphism h such that $h : (M_1, P_1,) \to (M_3, P_3)$.

Therefore, the morphisms commute.

Therefore, *Sto* is a category.

□

The proof can be shown by the following commutative diagram:

$$\begin{tikzcd}
(M_1,P_1) \arrow[d, "h"'] \arrow[dr, "f"] & \\
(M_3,P_3) & (M_2,P_2) \arrow[l, "g"]
\end{tikzcd}$$

Figure 2: A commutative diagram for the category *Sto*.

Example 4.5. *Given a column stochastic matrix M_1 and a partition P, find M_2.*

$M_1 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$

$S = \{1, 2, 3\}$

$P = \{\{1, 3\}, \{2\}\}$

The stationary distribution of $M_1 = \begin{bmatrix} \frac{2}{5} \\ \frac{1}{5} \\ \frac{2}{5} \end{bmatrix}$

$C = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$

$D = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \\ \frac{1}{2} & 0 \end{bmatrix}$

$D_{1,2}$

$$\begin{aligned}
&= Prob_{\{1,3\}\to\{2\}} \\
&= \frac{\pi_1}{\pi_1 + \pi_2}(\frac{1}{2}) + \frac{\pi_2}{\pi_1 + \pi_2}(\frac{1}{2}) \\
&= \frac{\frac{2}{5}}{\frac{2}{5} + \frac{2}{5}} + \frac{\frac{2}{5}}{\frac{2}{5} + \frac{2}{5}} \\
&= \frac{1}{2}
\end{aligned}$$

$CM_1D = \begin{bmatrix} 0 & \frac{1}{4} \\ 1 & \frac{3}{4} \end{bmatrix} = M_2$

5 Functors

Definition 5.1. *A* functor *F from a category C to a category D assigns an object in C to an object in D and assigns each morphism $f : x \to y$ in C to a morphism $F(f) : F(x) \to F(y)$ in D. The following properties must also hold:*

- $F(1_x) = 1_{F(x)}\ \forall x \in C$
- $F(g \circ f) = F(g) \circ F(f)\ \forall$ *morphisms* $f : x \to y$ *and* $g : y \to z$

So functors must also preserve identity morphisms and composition of morphisms. We hope to create a functor that will be a natural way to transform the category *Sym* to the category *Sto* and to also create a functor that will allow us to transform *Sto* back to *Sym.*

We will define the functor $F : Sym \to Sto$ as follows:

Proposition 5.2. *Let* Sym *be the category of symmetric matrices and let* Sto *be the category of stochastic matrices as defined above. Then $F : Sym \to Sto$ is a functor such that*

$$F(D_1, P_1) = (M_1, P_1) \tag{4}$$

$$F(f) = L \tag{5}$$

The symmetric distance matrix D is transformed into the column stochastic matrix M by the following algorithm:

$$\frac{\frac{1}{D_{ij}}}{\sum\limits_j \frac{1}{D_{ij}}} \tag{6}$$

where D_{ij} is the ij^{th} entry of the symmetric distance matrix D. The formula takes the reciprocal distance of D_{ij} and divides this by the sum of the reciprocal distances in each column of the matrix. So, the formula essentially normalizes each column in the matrix, transforming the distance matrix D into a stochastic matrix M. The functor F maps every symmetric matrix in Sto with non-negative real entries to a stochastic matrix in Sym and maps every morphism $L \in Sym$ to a morphism $L \in Sto$.

Proof. In order to prove that F is a functor, we must show that F preserves identity morphisms and composition of morphisms, as well as commutativity between categories.

identity: Let $(D, P) \in Sym$ be given and let $1_{(D,P)}$ be the identity morphism in *Sym* corresponding to (D, P).

Also, let $1_{F(D,P)}$ be the identity morphism in *Sto* corresponding to $F(D, P)$.

We need to show that $F(1_{(D,P)}) = 1_{F(D,P)}$. In the category *Sym*, the identity morphism $1_{(D,P)}$ creates a trivial partition where each subset of the partition contains only one element. Similarly in the category *Sto*, the identity morphism $1_{F(D,P)}$ also creates a trivial partition where each subset contains only one element.

So the functor F maps the morphism

$1_{(D,P)} : (D, P) \to (D, P)$ in *Sym* to $F(1_{(D,P)}) : F(D, P) \to F(D, P)$ in *Sto*.

Therfore, $F(1_x) = 1_{F(x)}$.
Therfore, F preserves identity morphisms.

composition: Let f and g be morphisms in the category Sym such that
$f : (D_1, P_1) \to (D_2, P_2)$ and $g : (D_2, P_2) \to (D_3, P_3)$.
Also, let $F(f)$ and $F(g)$ be morphisms in the category Sto such that
$F(f) : F(D_1, P_1) \to F(D_2, P_2)$ and $F(g) : F(D_2, P_2) \to F(D_3, P_3)$.
We need to show that $F(g \circ f) = F(g) \circ F(f)$.
$F(g \circ f) = F(g(f(D_1, P_1))) = F(g(D_2, P_2)) = F(D_3, P_3)$ and
$F(g) \circ F(f) = F(g(F(f(D_1, P_1)))) = F(g(F(D_2, P_2))) = F(D_3, P_3)$
Therfore F preserves composition of morphisms.

commutativity: Let Sym and Sto be the categories as previously defined.
Let (D_1, P_1) be an object in the category Sym, where D_1 is a $n \times n$ symmetric distance matrix and P_1 is a partition such that $P_1 = \{\{1\}, \{2\}, ..., \{n\}\}$.
Also, let (D_2, P_2) be an object in the category such that $f : (D_1, P_1) \to (D_2, P_2)$.
Let D_2 be a $k \times k$ matrix and let P_2 be a partition of P_1 such that $P_2 = \{\phi_1, \phi_2, ...\phi_k\}$.
Let (M_1, P_1) be an object in the category Sto, where M_1 is a $n \times n$ column stochastic matrix.
Also, let (M_2, P_2) be an object in the category such that $L : (M_1, P_1) \to (M_2, P_2)$, with M_2 being a $k \times k$ lumped matrix.
$F \circ f = L \circ F$
Therfore, F is commutative.
Thefore, F is a functor that maps from the category Sym to the category Sto.

□

The proof that F is a functor can be shown through the following commutative diagram:

$$\begin{tikzcd}
(D_1,P_1) \arrow[r, "F"] \arrow[d, "f"'] & (M_1,P_1) \arrow[d, "L"] \\
(D_2,P_2) \arrow[r, "F"'] & (M_2,P_2)
\end{tikzcd}$$

Figure 3: A commutative diagram for the for the functor F that maps from the category Sym to the category Sto.

In order to completely justify that EvoHop is a natural way to transform protein distances to phage distances we need to be able to move in the opposite direction between categories. That is, we need to define a functor G such that $G : Sto \to Sym$. G will be defined as follows:

Proposition 5.3. *Let Sym and Sto be the categories as previously defined. Let v be a matrix where the diagonal entries are made up of the stationary distribution of a stochastic matrix.*

$$v = \begin{bmatrix} \pi_1 & 0 & \dots & 0 \\ 0 & \pi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \pi_k \end{bmatrix}$$

Let $S = M_2 v$. Let $X = \frac{|\phi_i| \times |\phi_j|}{S_{ij}}$

$G : Sto \rightarrow Sym$ is a functor such that

$$G(M_1, P_1) = (D_1, P_1) \tag{7}$$

$$G(L) = (f) \tag{8}$$

The stochastic matrix M_1 is transformed into the symmetric distance matrix D_1 by the following algorithm:

$$\frac{X}{\sum_{i \in D} \sum_{j \in D} \frac{1}{D_{ij}}} \tag{9}$$

Essentially, this functor G takes a lumped stochastic matrix and transforms it into a symmetric distance matrix.

Proof. In order to prove that G is a functor, we need to show that G preserves indentity morphisms and composition of morphisms and also that G commutes within the category.

identity: Let $(M, P) \in Sto$ be given and let $1_{M,P}$ be the identity morphism in Sto corresponding to (M, P). Also, let $1_{G(M,P)}$ be the identity morphism in Sym corresponding to $G(M, P)$.

We need to show that $G(1_{(M,P)}) = 1_{G(M,P)}$.

Therefore, G preserves identity morphisms.

associativity: Let L and N be morphisms in the category Sto such that
$L : (M_1, P_1) \rightarrow (M_2, P_2)$ and $N : (M_2, P_2) \rightarrow (M_3, P_3)$.
Also, let $G(L)$ and $G(N)$ be morphisms in the category Sym such that
$G(L) : G(M_1, P_1) \rightarrow G(M_2, P_2)$ and $G(N) : G(M_2, P_2) \rightarrow G(M_3, P_3)$.
We need to show that $G(N \circ L) = G(N) \circ G(L)$.
$G(N \circ L) = G(N(G(M_1, P_1))) = G(N(M_2, P_2)) = G(M_3, P_3)$
$G(N) \circ G(L) = G(N(G(L(M_1, P_1)))) = G(N(G(M_2, P_2))) = G(M_3, P_3)$
Therefore G preserves composition of morphisms.

commutativity: Let Sym and Sto be our categories as previously defined. Let (M_1, P_1) be an object in the category Sto where M_1 is a $n \times n$ column stochastic matrix and P_1 is a aprtition such that $P_1 = \{\{1\}, \{2\}, ...\{n\}\}$.

Also, let (D_2, P_2) be an object in the category such that $L : (M_1, P_1) \rightarrow (M_2, P_2)$.

Let M_2 be a $k \times k$ matrix and let P_2 be a partition of P_1 such that $P_2 = \{\phi_1, \phi_2, ..., \phi_k\}$.

Let (D_1, P_1) be an object in the category Sym where D_1 is a $n \times n$ symmetric distance matrix.

Also, let (M_2, P_2) be an object in the category such that $f : (D_1, P_1) \rightarrow (D_2, P_2)$, with D_2 being a $k \times k$ matrix.

$G \circ L = f \circ G$

Therefore, G is commutative.

Therefore, G is a functor that maps from the category Sto to the category Sym.

□

The proof that G is a functor can be shown by the following commutative diagram:

$$\begin{tikzcd}
(M_1,P_1) \arrow[r, "G"] \arrow[d, "L"'] & (D_1,P_1) \arrow[d, "f"] \\
(M_2,P_2) \arrow[r, "G"'] & (D_2,P_2)
\end{tikzcd}$$

Figure 4: A commutative diagram for the for the functor G that maps from the category Sto to the category Sym.

6 Natural Transformation

EvoHop can be shown to be a natural transformation by the following commutative diagram:

$$\begin{tikzcd}
Sym \arrow[r, "F"] \arrow[d, "f"'] & Sto \arrow[d, "L"] \\
Sym & Sto \arrow[l, "G"]
\end{tikzcd}$$

Figure 5: The natural transformation justifying EvoHop

Since one can achieve the desired result in one step rather than three, it begs the question as to why it is necessary to map between categories, lump, and then map back to the original category. This process is needed in order to calculate the stationary distribution of the reversible Markov chain which will then tell us the importance ranking of each phage protein.

7 Discussion

Within reversible Markov chains, there exists a well-established, natural way to lump together elements in the chain. This is what acted as a morphism in the category *Sto* [3], [2]. However, there does not already exist a natural way to lump together elements in a distance graph, represented by a symmetric matrix. A number of different methods exist in transforming a symmetric matrix to a lumped symmetric matrix. One methods is to simply sum the distances between subsets of the partition. Another method is the average the distances between subsets of the partition, using an arithmetic, geometric, or harmonic mean. Initially, the harmonic mean was thought to be the natural way to lump, since calculating the reciprocal of each distance would produce rates [4]. Each method was attempted and calculated using various examples. However, we sought the precise moethod that would correspond with thelumping algorithm of reversible Markov chains. The correct lumping method that corresponded to lumping of reversible Markov chains was determined to be taking the harmonic mean of the distances. The $|\phi_i| \times |\phi_j|$ factor was also not known at the beginning of the research. This factor was determined through example calculations.

8 Future Problems

There are still problems that need additional work on this probject.

- One problem is that we need to incorporate a scaling factor into set of morphisms in category *Sym*, such as

$$(D_2)_{ij} = \frac{|\phi_i| \times |\phi_j|}{\sum\limits_{\ell \in \phi_i, m \in \phi_j} \frac{1}{D_{\ell m}}} x \tag{10}$$

  where x is a real number. This scaling factor x allows you to transform symmetric distance matrices while preserving the symmetry of the matrix.

- The proofs of the categories also need some additional work. It was determined very late in the project that the associativity and identity of the morphisms within the category needed to be formally proven.

- The functor G that maps from the category *Sto* to the category *Sym* also needs to be worked through with more example calculations.

- A formal proof that EvoHop is indeed a natural transformation needs to be stll be written.

References

[1] Annalinda Arroyo, K. Mecadon, R. Edwards, F. Rohwer, P. Salamon
"Using GoogleRanks for Phage Phylogeny"

[2] C.J. Burke, M.Rosenblatt, 1958 "A Markovian Function of a finite Markov Chain" *Ann. Math Statist.* 29:1112-1122

[3] Kemeny, John G. Snell, J. Laurie 1960.*Finite Markov Chains*. New York.

[4] B. Andersen, K.H. Hoffman, K. Mosegaard, J. Nulton, J.M. Pedersen, P. Salamon, 1988 "On Lumped Models for Thermodynamic Properties of Simulated Annealing Problems." *J. Phys* 49:1485

[5] Saunders Mac Lane 1998. *Categories for the Working Mathematician*. Springer.

[6] David L. Swofford, G. Olsen, P. Waddell, D. Hillis 1996. *Phylogenetic Inference*. 407-514

[7] Wikipedia 2007. *Category Theory*