Evolutionary Protein Distances to Phage Evolution Model

Andrew Detzel (Grad Student Assistant for Phage Research Project)

SDSU REUT Summer 2007

1 Introduction

During the eight weeks between June 4th and July 27th, the 2007 SDSU REUT phage research group investigated mathematics pertaining to viruses that predate on bacteria called phages. The aim was to contribute original research that can be applied towards phage analysis software called Phage Communites by Contig Spectrum (a.k.a. PHACCS). This online tool has the capacity of modeling the structure and estimating the diversity of uncultured viral communities [1]. The term Phylogeny refers to evolutionary history. Biologist are interested in the family tree like phylogeny organization of organisms.[2] Underneath the larger goal, the personal goal was to produce a probabilistic model to characterize the evolutionary behavior of the phage evolution. Others had previously attempted this though without sufficient justification. The first efforts involved numerous failed attempts employing combinatorial and category theoretical arguments to justify previous models Ultimately, the model was constructed using an analogy argument solidified by category theory[3].

2 Mathematical Background

The following is a lecture partially given and maintained throughout the program by Andrew Detzel to provide the operational Probability and Markov Chains knowledge for the group based on [4, 5, 6]. It and the brief category theory intro following it furnish sufficient information to allow a typical mathematics undergraduate to understand the remainder of the document.

2.1 Probability

Let Ω be a set and let σ be a collection of subsets of Ω. We call σ a $\sigma-algebra$ if:

a) $A \in \sigma \Rightarrow A^c \in \sigma$, and

b) $A_1, A_2, A_3, ... \in \sigma \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \sigma$.

In English, this simply means that a $\sigma - algebra$ is a set of subsets of Ω that is closed (like in Algebra) under the operations of union and compli-

ment. In probability, events are represented by subsets of a set of outcomes, and the set of these events is the $\sigma - algebra$. An ordered pair of a set with one of its $\sigma - algebras$ like (Ω, σ) is called a *measurable space*. A measure μ on a set Ω is a function $\mu : \sigma \to [0, \infty)$ such that:

a)$\mu(A) \geq \mu(\emptyset) = 0$, $\forall A \subseteq \Omega$, and

b) if $\{A_i\}_{i=1}^{\infty}$ is a countable collection of disjoint sets in σ then

$\mu(\bigcup_{i \in I} A_i) = \sum_{i \in I} \mu(A_i)$.

A probability function $\mathbb{P}$ on Ω is a measure on Ω such that $\mathbb{P}(\Omega) = 1$. In this case, we call Ω the *Outcome Space*. For example, heads, tails is the outcome space for a coin flip and the $\sigma - algebra$ would be $\{\emptyset, \{heads\}, \{tails\}, \{heads, tails\}\}$ the probabilities would be defined as follows: $\mathbb{P}(heads) = \mathbb{P}(tails) = 1/2$, $\mathbb{P}(heads, tails) = \mathbb{P}(\emptyset) = 0$, and $\mathbb{P}(heads \bigcup tails) = \mathbb{P}(\Omega) = 1$.

If a probability distribution is defined on a continuum like $\mathbb{R}$ then it is called *continuous* and if the outcome space is discrete like $\mathbb{Z}$, it is called *discrete*. In the continuous case, the probability function is defined by a *density function*. So, if f is the density function of $\mathbb{P}$, then $\mathbb{P}(A) = \int_A f d\mu$.

Proposition 2.1. $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$

Proof. $A \bigcup A^c = \Omega$ and A is disjoint from A^c so $1 = \mathbb{P}(\Omega) = \mathbb{P}(A \bigcup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$ and thus $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$. □

Given outcome space a *Random Variable* -denoted by a capital letter- is a function on the outcome space. For example, if we define C by $C(heads) = 1$ and $C(tails) = 0$, then C is a random variable. Consider the coin-flip example. Observe it would make sense to say $\mathbb{P}(C = 1) = \mathbb{P}(heads)$ since $heads = C^{-1}(1)$. So, given a probability function $\mathbb{P}$, we can assume define probabilities of values of a random variable X as follows:

$\mathbb{P}(X = x) = \mathbb{P}(X^{-1}(x))$.

Similarly for a set B of values that X can take, we can define

$\mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B))$.

With definition we can say that the random variable X has the distribution P. Henceforth Ω will be an outcome space, $\mathbb{P}$ will be a distribution on that space and X will be a random variable on Ω with distribution $\mathbb{P}$.

2.1.1 Expectation, Means and Variance

Now that we have axiomatically defined probability, several other quantities interest us. If $\mathbb{P}$ is a discrete distribution then define $\mathbb{E}(X) = \sum_{x \in \Omega} x\mathbb{P}(X = x)$. If $\mathbb{P}$ is a continuous distribution with density function f, then we define $\mathbb{E}(X) = \int_{\Omega} xf(x)dx$. In either case we call $\mathbb{E}(X)$ the *Expected Value* of X. The *Mean* of a distribution is the expected value of a random variable with that distribution. Now, we call the quantity $\sigma^2 := \mathbb{E}((X - \mathbb{E}(X))^2)$ the *Variance* of X or $Var(X)$ and we call $\sqrt{(}\sigma^2)$ the *Standard Deviation* of X or $SD(X)$.

2.2 Central Limit Theorem

Theorem 2.2. *Given a sequence of independent and identically distributed random variables $X_1, X_2, ...$ with finite variance σ^2 and expected value μ, then the distribution of $(\sum_{i=1}^{n} X_i)/n$ converges to a normal distribution with mean μ and variance σ^2.*

Proof. This proof requires the inverse Fourier transform which is beyond the scope of this document. □

2.3 Markov Chains

A *Stochastic Process* $(X_t)_{t \geq 0}$ is a sequence of random variables indexed by time. If the time-index is discrete we would denote the process by $(X_n)_{n \geq 0}$. A Stochastic Process $(X_n)_{n \geq 0}$ is said to be a Markov Chain if $\mathbb{P}(X_n = i_n | X_0 = i_0, X_1 = i_1, ..., X_{n-1} = i_{n-1}) = \mathbb{P}(X_n = i_n | X_{n-1} = i_{n-1})$. We call the set of values that the Markov chain takes the *State Space* and we will denote it by S. Unless explicitly stated, for the remainder of this paper, $(X_n)_{n \geq 0}$ will be a Markov Chain with state space S. We will also refer to a Markov Chain with the term *random walk*.

2.3.1 Transition Probabilities

The probability $p_{ij} := \mathbb{P}(X_{n+1} = j | X_n = i)$ is called the 1-step *Transition Probability*. The matrix P whose $i-jth$ entry is p_{ij} is called the *Transition Matrix* of $(X_n)_{n\geq 0}$. As it turns out (the proof will follow) the $i-jth$ entry of the matrix P^n (denoted by $p_{ij}^{(n)}$) is the probability $\mathbb{P}(X_{m+n} = j | X_m = i)$. We call the p_{ij}^n's the *n-step Transition Probabilities*. These turn out to not be hard to calculate sometimes.

In the case where the transition probabilities are independent of time, the problem reduces to finding the nth power of the transition matrix but as we ALL remember from Linear Algebra, we can diagonalize our matrix then raise it to a high power so that $P^{(n)} = UD^{(n)}U^{-1}$ where D is the diagonal matrix of eigenvalues of P. Thus after a little bit of thought we can conclude that $\exists$ constants $a_1, a_2, ...a_k$ such that $p_{ij}^n = \sum_{j=1}^k a_j \lambda_j^n$ where λ_j, $i = 1, 2, ..., k$ are the eigenvalues of P.

Example. Let $P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \end{pmatrix}$

The eigenvalues of P are: $1, i/2, -i/2$, so $p_{ij}^n = a + b(i/2)^n + c(-i/2)^n$ for some constants a, b, c.

Theorem 2.3. *$(X_n)_{o\leq n\leq N}$ is Markov iff for all $i_1, i_2, ..., i_N \in I$:*

$\mathbb{P}(X_0 = i_0, X_1 = i_1, ..., X_N = i_N) = \lambda_{i_o} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{N-1}} p_{i_N}$, *where λ is the distribution of X_0.*

Theorem 2.4. *Let $(X_n)_{n\geq 0}$ be Markov with transition matrix P. Then, given that $X_m = i$, $(X_{m+n})_{n\geq 0}$ is Markov with transition Matrix P and is independent of $X_0, X_1, ..., X_m$.*

Theorem 2.5. *Let $(X_n)_{n\geq 0}$ be Markov with transition matrix P and suppose X_0 has distribution λ. Then:*

(i) $\lambda P^{(n)}$ is the vector whose ith entry is the probability of being in state i at time n, and

(ii) $\mathbb{P}(X_n = j | X_0 = i) = p_{ij}$.

Proof. (i) $\mathbb{P}(X_n = j) = \sum_{i_0,i_1,...,i_{n-1}} \mathbb{P}(X_0 = i_0, X_1 = i_1, ..., X_n = j) = \sum_{i_0,i_1,...,i_{n-1}} \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n} = (\lambda P^{(n)})_j$

(ii) Conditioning on $X_0 = i$ just means (by the previous theorem) $\lambda_i = 1$ and $\lambda_j = 0$ for $i \neq j$. Thus, by part (i) $\mathbb{P}(X_n = j | X_0 = i) =$
$\sum_{i_0,i_1,...,i_{n-1}} \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n} = \sum_{i_1,...,i_{n-1}} \lambda_i p_{i i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n} = p_{ij}^{(n)}$. □

Corollary 2.6. ***Corollary to Perron-Frobenius Theorem***
Given a stochastic matrix P such that all entries of $P^{(n)}$ are strictly positive for some n, then:

a) P has 1 as an eigenvalue,

b) all other eigenvalues of P have absolute value less than 1, and

c) P has a unique eigenvector of eigenvalue 1 with all positive entries. (If we scale it so that the entry sum is 1, then this is our stationary distribution).

2.3.2 Reducibility

One natural question that arises in Markov chains is starting from state i, what states do I have a chance of getting to. We say that state i leads to state j and write $i \to j$ if $\exists n \in \mathbb{N} \ni p_{ij}^{(n)} > 0$. We then say state i *communicates* with state j and write $i \leftrightarrow j$ if $i \to j$ and $j \to i$. The first thing one would see is that "communication" is an *equivalence* relation. So then the natural thing to discuss is *communication classes*. These classes are subsets of the state space which are separated from the other parts of the graph. If your transition matrix is associated with a single class, it is called *irreducible*. Otherwise, it is irreducible. Consider the transition matrix

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

As one can see the Markov Chain represented by P is on two triangles $1, 2, 3$ and $4, 5, 6$ so if the Chain starts at $1, 2$ or 3 it will never reach $4, 5$ or 6 and vice-versa. This is where the shadow graphs come in. For our phage-matrices we want our matrix to be irreducible so we would give minute ϵ sized transition probabilities to make out matrix irreducible, for example, the matrix P above would become:

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & \epsilon & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \epsilon & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

so that now there is a connection between $1, 2, 3$ and $4, 5, 6$.

2.3.3 Hitting Probabilities and Expected Hitting Time

One natural question to ask is, starting from state i what is the probability that a chain hits state j? Let's define $h_i^j = \mathbb{P}(\text{Hit } j | X_0 = i)$ and $k_i^j = \mathbb{E}(\text{Time to hit } j | X_0 = i)$. We call h_i^j a *hitting probability* and k_i^j an *expected hitting time*.

Example.
Consider a chain with matrix

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

So, starting from state 2 what is the probability that the chain gets to state 4 and how long would you expect it to take to reach 1 or 4?

Consider h_2^4 and k_2^4. Let's take a step. In one step, starting from state 2 we must move to state 1 with probability $1/2$ or state 3 with probability $1/2$. Thus, $h_2^4 = p_{21}h_2^1 + p_{23}h_2^3 = 0 + 1/2h_2^3$ and $k_2^4 = 1 + p_{21}k_1^4 + p_{23}^4 = \infty$ (since starting from 1, one will never get to 4). The process we used here generalizes to give us the following Theorems:

Theorem 2.7. *The vector of hitting probabilities $h = (h_i^j, i \in I)$ is the minimal non-negative solution to the system:*

$h_j^j = 1$
$h_i^j = \sum_{k \in I} p_{ik} h_k^j$

Theorem 2.8. *Similarly, the vector of mean hitting times $k^j = (k_i^j, i \in I)$ is the minimal non-negative solution to the system:*

$k_j^j = 0$
$k_i^j = 1 + \sum_k p_{ik} h_k^j$

2.3.4 Invariant Distributions

Recall, if X_0 has distribution λ then the vector whose ith entry is the probability of being in state i at time n is $\lambda P^{(n)}$. Our last main question then is, given a transition matrix P, does there exist a distribution π such that $\pi P = \pi$?

2.3.5 Rich Example

Consider a Markov Chain with transition matrix:

$$P = \begin{pmatrix} 1/4 & 1/2 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Consider the asymptotic behavior of the figure and determine transition probabilities from each state to each other.

2.4 Continuous Time Markov Processes

Let I be a countable set. A *Q-Matrix* on I is a matrix $Q = (q_{ij} : i, j \in I$ such that

(i) $q_{ij} \geq 0$ for all $i \neq j$.

(ii) $\sum_{j \in I} q_{ij} < \infty$ and $q_{ii} = -\sum_{j \in I} q_{ij}$

The way to construct a Q matrix is to make q_{ij} the rate at which a process goes from state i to state j for off diagonal entries and define $q_{ii} = -\sum_{j \in I} q_{ij}$, sort of a leaving rate of state i. Let $(X_t)_{t \geq 0}$ be the continuous time stochastic process which has these rates. Now make a family of matrices $P(t)$ where $p_{ij} := \mathbb{P}(X_t = j | X_0 = i)$. Observe these are stochastic matrices. In this case we call $(X_t)_{t \geq 0}$ a *Markov Process* with Now, as it turns out, if we make our Q-matrix with these rates then, $(X_t)_{t \geq 0}$ is a *Markov Process* and $P(t) = e^{(tQ)}$ defines a continuous time transition matrix for $(X_t)_{t \geq 0}$, i.e. $p_{ij}(t) = \mathbb{P}(X_t = j | X_0 = i)$. ***Note that a continuous time process is Markov if $\mathbb{P}(X_{t_{n+1}} = j | X_{t_0} = i_0, X_{t_1} = i_1, ..., X_{t_n} = i) = \mathbb{P}(X_{t_{n+1}} = j | X_{t_n} = i)$

2.4.1 Invariant Distributions

If $(X_t)_{t \geq 0}$ is a Markov Process with Q-matrix Q and transition matrix P, then a distribution π is an *Invariant Distribution* of $(X_t)_{t \geq 0}$ if $\pi Q = 0$.

2.5 Categories and Functors

All of the following comes from a single source.[7]

Definition 2.9. *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 a binary operation $\circ$ such that if $f : a \to b, g : b \to c$, and $h : c \to d$ are morphisms then:*

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

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

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

Definition 2.10. *A* Functor *$F : C \to D$ is a map from a category C into a category D.*

Definition 2.11. *A Functor $F : C \to D$ is said to be* covariant *if for each morphism $f : x \to y$ $F(f) : F(x) \to F(y)$.*

3 Final Research Paper

The culmination of Andrew Detzel's research during the 2007 SDSU math REUT is given by the following research paper. It has received much and invaluable direction by Peter Salamon and is targeted to be a brief paper in a journal on the order of *Physica A*.

An Electrical Network Phage Evolution Model Based on Protein Distances Andrew Detzel July 27, 2007

Abstract

The aim of this paper is to present a model for phage evolutionary information flow analogous to a random walk on an electrical network by considering evolutionary distance like electrical resistance. In the process, we convert a protein distance network into an inter-phage distance network then assign a transition matrix for a random walk to the distance network. The set of distance networks forms a category with the morphisms defined by the map we use to convert protein distances into phage distances as does the set of transition matrices and partitions with morphisms being the lumping of matrices with respect to the partitions. The conversion from phage distance network to transition matrix is a functor which is covariant with respect to the morphisms defined for the distance networks. With the model in place, we can employ centrality measures which are important for phylogeny.

Nucleotides are the building blocks of DNA. Three nucleotides code for amino acids which compose proteins. Empirical data can roughly determine inter-amino acid evolution rates which can in turn be used to create a Markov Model simulating amino acid and protein evolution as done by the likes of Simon Whelan and Nick Goldman [8]. In the context of phage phylogeny, a similar model for species evolution is desired. These evolutionary distances are analogous ro resistance in an *electrical network*, that being a network in which the edges are weighted with real-valued resistances[9].

Consider a protein distance network D^* (in matrix form), suppose it is $m \times m$. Each of these proteins corresponds to a phage, say there are n phages. Thus, we form a partition $\phi_1, \phi_2, ...\phi_n$ of the network by letting each ϕ_i be the set of proteins contained in phage i. We now wish to convert D^* into a phage distance network D so we define the nodes of D to be $\phi_1, \phi_2..., \phi_n$ and assign distances to the edges as follows:

$$d_{ij} = \frac{|\phi_i||\phi_j|}{\sum_{k\in\phi_i, l\in\phi_j} \frac{1}{d^*_{kl}}}. \tag{3}$$

This matrix yields a (simple) distance network where the weights are the harmonic means of all the pairwise protein distances of the phages. We use the harmonic means as they turn out to be the morphisms which makes our transition from distance networks to random walk transition matrices a covariant functor. For the purposes of our evolution model, we now change forms of our evolutionary electrical network from distance/resistance to closeness/conductance as [9] provides our random walk model for this form of the network. The conductance of an edge ij is given by:

$$c_{ij} = \begin{cases} \frac{1}{d_{ij}}, & ij \in Edges(D) \\ 0, & otherwise \end{cases} \tag{4}$$

We note though, that in phages, multiple phages sharing a single protein is so rare [10] that we ignore the possibility making our conductance network a simple one. It follows that our conductance then aligns exactly with a parallel notion of closeness in a network. There are many ways of measuring closeness though the most obvious in this case and a very well established one [11] is given by the closeness c_{ij} of two nodes (phages) being:

$$c_{ij} = \begin{cases} \frac{1}{d_{ij}}, & i \neq j \\ 0, & i = j \end{cases} \tag{5}$$

Since the closeness and conductance are the same in this case, we denote both by c_{ij}.

The next step is to consider the evolutionary flow along our evolutionary closeness/conductance network. We apply the conductance random walk model as defined in [9]. That is, we define a random walk $(\{X_t\}, P)$ on C by:

$$p_{ij} := \frac{c_{ij}}{\sum_j c_{ij}}. \tag{6}$$

The map defined above from distances (and partitions) to transition matrices is a covariant functor which preserves the morphisms defined by equation (1) [3]. We notice that the diagonal entries of the transition matrix are zero since the phages don't share proteins. This allows us to take the interpretation of our model to be the jump process underlying the real-time evolution. In other words, the time this model runs on is on the count of evolutionary transitions as opposed to real-time.

Of immediate interest to phylogeny is influence flow and importance rankings of phages. The most obvious quantity would be the random walk's equilibrium distribution. Given a large enough population and enough time elapsed, the equilibrium distribution yields about the relative proportion of phage species and thus a ranking of importance. A value C_i called random walk centrality quantifies how central phage i is with respect to its receptivity of evolutionary flow [12]. This would provide a ranking of which phages have the greatest evolutionary advantage as opposed to just some frequency of phages that is converged to over time (equilibrium distribution).

References

[1] Peter Salamon. http://eli.sdsu.edu/phagewiki/2. Reu introductory talk: Metagenomics in phage ecology.

[2] G. J. Olsen P. J. Waddell Swofford, D. L. and D. M. Hillis. *Molecular Systematics (2nd ed.)*. Sinauer Associates, Sunderland, Massachusetts, 1996.

[3] David Aaby. Using category theory to justify evohop.

[4] James Norris. *Markov Chains*. Cambridge University Press, Cambridge, United Kingdom, 1997.

[5] Gregory Lawler. *Introduction to Stochastic Processes*. Chapman and Hall, Boca Raton, Florida, 1995.

[6] Rick Durrett. *Probability: Theory and Examples*. Thomson, Belmont, California, 2005.

[7] Saunders Mac Lane. *Categories for the Working Mathematician*. Springer-Verlag, New York, New York, 1971.

[8] Simon Whelan and Nick Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, 18:691–699, 2001.

[9] Béla Bollobás. *Modern Graph Theory*. Springer, New York, NY, 1998.

[10] Peter Salamon.

[11] Chavdar Dangalchev. Residual closeness in networks. *Physica*, 2006.

[12] Jae Dong Noh and Heiko Rieger. Residual closeness in networks. *Physica*, 2006.