Towards a Phylogeny of Bacteriophage via Protein Importance Ranking

Kevin Manley

July 27, 2007

Abstract

Protein importance ranking is a step in a process that may yield a phylogeny for bacteriophage. Irazoque, et al.[1] began with protein distances and proceeded to protein transition probabilities, phage transition probabilities, and phage distances respectively via a method that they call Evohop. Here I examine the potential for an Evohop that begins with phage protein transition rates. In the Evohop method, protein importance ranks can be found from the stationary distribution of the protein transition probabilities. Research results for the method examined in this paper have not yet been attained.

1 Introduction

In the past few years, there has been growing interest in bacteriophage. Bacteriophage are bacteria eating virus that predominate in the ocean but can be found anywhere that bacteria can be found. Recent research has been directed towards modelling bacteriophage communities[7], and inferring phage phylogeny[8]. The growth of interest in bacteriophage may be due to the potential that this bacteria eating virus has for fight disease, the potential phage have for processing carbon, or simply because of the large amount of this ubiquitous organism that can be found nearly anywhere that one can find bacteria.

The growth of interest in bacteriophage coincides with recent advances in molecular biological technologies. Since the 1980s the number of sequenced genomes that are catalogued at the genome bank that is maintained by the National Center for Biological Information has grown exponentially. From 2002 to 2005 alone the number of sequenced genomes in the GenBank database has risen from 24 million to 52 million[3]. During the same period

the number of known phage grew from around 100 to approximately 500. Five years ago, the group that developed Phage Communities from Contig Spectra (PHACCS) used just over 1000 phage DNA fragments in their study[7]. The same goup today, could easily find themselves working with hundreds of thousands of fragments of bacteriophage DNA. As the database of information grows, biologists, mathematicians, and computer scientists will need to work together to develop the most computationally efficient methods of analysis possible.

Phylogeny attempts to describe, statistically, the similarity or difference between groups of species with an evolutionary tree. To the extent that computing power is not an issue in the development of species phylogenies, optimality criteria-based methods are probably most desirable. Optimality criteria methods, like maximum parsimony and maximum likelihood, require both an algorithm for find an evolutionary tree as well as a criteria for choosing the best of all possible trees. Distance-based methods of phylogeny, on the other hand, require only an algorithm, which proceeds directely to that tree which best represents the distances between species. All methods for determining the phylogeny of a group of species seek to infer an evolutionary history based upon a current state, where the current state may be identified by morphological data, DNA sequence data, or protein sequence data. In this paper, protein sequence data is used. Regardless of data being used, the number of possible trees for T taxa, or species, is

$$B(T) = \prod_{i=3}^{T}(2i-5).$$

Clearly the determination of an evolutionary tree for even several species is computationally costly. In this computational environment, particularly where phage are concerned, the large amount of data available is more suited to distance methods[6, p. 175].

Distance methods, such as neighbor-joining and other clustering methods move directly towards an evolutionary tree based upon algorithms that are typically easier to program and faster than the optimal criteria methods that require both an algorithm and a criteria. The distance between two species can be thought of as the branch length between the species in an evolutionary tree[6, p. 147]. This distance is a function of time and expected amount of evolution. As such, distance methods require both some measure of the extent to which the DNA or protein sequences are similar as well as an evolutionary model, also known as a transition or Q matrix. Protein distances can be calculated using any of a number of bioinformatics tools. General evolutionary models are available, but in this paper the transition matrix is calculated based upon available bacteriophage-specific data.

2 Importance Ranking

2.1 Protein Distance-Based Evohop

Evohop is a method for deriving the distance between species from the distance between their proteins[1]. The method suggests that bacteriophage protein distances, for instance, may be used to find bacteriophage transition probabilities. The tranistion probability matrix is a square matrix where each entry represents the probability of transition from one protein to another, with many of the entries zero, because often the pair of proteins is so different that there is virtually no possibility of one protein evolving into the other. Note that protein to protein evolution is merely a theoretical construct that is analagous here to a measure of how closely related two proteins are. This stands apart from any interest that one may or may not have in reverse evolution. The next step would be to lump the transition probabilities according to the bacteriophage in which they are found to find a bacteriophage transition probability matrix. Again, the bacteriophage transition probability matrix is a theoretical construct that is a measure of how closely pairs of phage are related. With the bacteriophage transition probability matrix we would proceed to a bacteriophage distance matrix that would be used to build an evolutionary tree using a distance-based phylogenetic method. In the work of Salamon, et.al.[1] and this paper, however, we follow the above steps only until we find the transition probability matrix, opting instead to rank the importance of bacteriophage proteins.

In the protein distance-based evohop method protein importance rankings are determined by converting the protein distance matrix, D, to a protein transition probability matrix with

$$P = \frac{\frac{1}{D_{ij}}}{\sum_n^{i=1} \frac{1}{D_{ij}}} + \epsilon G_{ij}.$$

where ϵG_{ij} is a coupling matrix that allows for the possibility of two proteins in the same bacteriophage evolving, one into the other. The protein transition probability matrix can be evaluated to find the protein importance ranks from the stationary distribution.

2.2 Transition Matrix-Based Evohop

An alternate method, which this researcher studied, for finding protein importance ranks begins with the amino acid transition rates for phage proteins. The transition rate for one amino acid, i, to another amino acid, j, is found

with

$$Q_{ij} = S_{ij} diag(\pi_i).$$

Q is the 20x20 amino acid to amino acid transition rate matrix. S_{ij} is a 20x20 amino acid to amino acid substitution matrix, which simply counts the number of occurrences of each amino acid pair. S_{ij} is symmetric and the diagonals of S_{ij} are zero. π_{ij} is the frequency of occurence of each amino acid in a selection of proteins.

The transition rate matrix is used to find a transition probablity matrix for the transition probability of one amino acid to another in a pair of proteins. Here the pairwise distance between two proteins can be found by optimizing the transition probability matrix P over t in

$$P(t) = e^{Qt}$$

where the time, t, is a measure of evolutionary time, or distance between the two proteins. The value of the function at the optimized t is a transition probability matrix for the amino acid transition probability. Note that the same amino acid transition rate matrix, Q, is determined for an entire family of proteins, but there is a different transition probability matrix, P, for every possible pair of proteins in the family.

The transition probability matrix is applied to each homologous site in a pair of aligned proteins to determine the relationship between a pair of proteins - the likelihood, L, that one protein will evolve into another.

$$L = \prod_{i}^{n} P_{1i} \bullet P_{2i}$$

where n is the number of aligned sites and i, each 1:n site, merely represents the extent to which two proteins are similar. We determine the likelihood for a pair of proteins in a family. This yields a matrix of likelihoods. A protein will not be evaluated for the likelihood of changing into itself, so the columns will be normalized after a zero is substituted for each diagonal entry. As with the protein distance-based evohop method, we find a stationary distribution matrix where the transitions reach an equilibrium and assign ranks to each protein based upon its likelihood in the the stationary distribution.

3 Results and Discussion

At the outset, let me say that I have no results to report at this time. To whatever extent the transition matrix-based evohop method is a valid method for finding protein importance ranks, the following discussion may have value.

Research outcomes were limited by time constraints, personal ability, and, perhap, failure to communicate effectively with other researchers and the project director. At some point, the discussion definitely takes a turn from what was done towards what could have been done.

The genome bank that is maintained by the National Center for Biotechnology Information has approximately 510 sequenced bacteriophage genomes and 27000 sequenced baceriophage proteins. The San Diego State University Bioinformatics Department provided approximately 500000 files that represented a relatively small portion of the nearly 27000^2 protein distances that could be calculated from available data. Each file contained two aligned protein sequences that were judged to be sufficiently close. The 500000 distances were further pared to approximately 120000 by taking only those pairs with distances less than 1.6 and e-values less than 10^{-4}.

The volume of the data was problematic, simply in terms of file management. The 500000 protein distances were delivered by the Bioinformatics Department in twelve different folders that were numbered 1, 1b, 1c, 1d, and 2 through 9. Each file was named with the pair of numbers that indexed the 27000 available proteins. File 324-23215, for instance, would contain bacteriophage protein number 324 and bacteriophage protein number 23215. Folders 2 through 9 included files such that the first digit in the first protein listed conincided with the number of the folder. Folder number 3, for instance, could have contained file 324-23215. Folders 1, 1b, 1c, and 1d, however, were not sorted in this manner. Each of these folders could contain a file with a first digit that was not a 1. I considered that the Matlab code that would be used to comb the 500000 files in search of the 120000 that met the criteria for distance and e-value would work faster if all of the files were sorted so that the first digit of the file matched the folder number. In execution, the process of placing the files in the proper folders, was time consuming. In the end, I displaced or deleted approximately half of the targeted 120000 files.

Having unsuccessfully sorted the files into the folders, the folders were searched and files were evaluated to determine the frequency of each amino acid in the approximately 60000 pairs of proteins. The files were also evaluated to determine the number of occurences of each possible amino acid pair, excluding any matching pairs. Note that in addition to the 20 amino acids, the counts included gaps that were inserted when the proteins were aligned, two types of ambiguous cases, and a category for unknown sites. The amino acids were placed in a 24x24 diagonal frequency matrix, π_i and the counts for each amino acid pair were placed in a symmetric 24x24 substitution matrix such that $s_{ij} + s_{ji} = S_{ij} = S_{ji}$. With the S diagonal zero, the diagonal of $Q = S_{ij} \bullet diag(\pi_i)$ was calculated to yield a row stochastic transition matrix.

The transition matrix, Q, identifies the transtion probability matrix for each pair of proteins when P is optimized over t in

$$P(t) = e^{Qt}$$

for each pair of proteins in the family of proteins that is being used to determine the distance between species. A fairly simple Matlab code that indexes each amino acid and the four additional characters mentioned above can be written so that the appropriate Q matrix entries can be retrieved to maximize the likelihood function

$$L_n = P_{AA_1} \bullet P_{AA_2} \bullet P_{AA_3} \bullet \ldots \bullet P_{AA_n}$$

where P_{AA_1} is the probability of transition from the first amino acid in protein 1 to first amino acid in protein 2, and P_{AA_2} is the probability of transition from the second amino acid in protein 1 to the second amino acid in protein 2, etc. Since $P = e^{Qt}$, we can say that

$$L_n = e^{Qt}_{AA_1} \bullet e^{Qt}_{AA_2} \bullet e^{Qt}_{AA_3} \bullet \ldots \bullet e^{Qt}_{AA_n}$$

where $e^{Qt}_{AA_1}$ is the entry in the transition probability matrix that correnspponds to the pair of amino acids found at the first homologous site, and $e^{Qt}_{AA_2}$ is the entry in the transition probability matrix that corresponds to the pair of amino acids found at the second homologous site, etc. The function L is optimized over t for one pair of proteins. The optimization over L is repeated for each protein in the family - in this case, approximately 500000 times, yielding a transition likelihood for each pair of aligned proteins. I developed a Matlab code, which I tested on two phage proteins, but I did not run the code on the entire database of aligned proteins and did not find importance ranks for phage proteins.

4 REUT Experience

I never felt like I was able to find stable footing during this research project. Each time that I felt I was coming to an understanding about what had already been previously achieved in this area of research, or what I was doing, or where I was headed, I would read something, ask a question of a graduate assistant, or ask a question of the project director that would reveal some fundamental misunderstanding of the problem at hand. Having some experience with research in areas outside the mathematical sciences, the most frustrating aspect of my research experience was a consistent inability to correlate

my activities with what I was reading in the literature. I read the papers that my project director assigned over and over again. I read the papers that those papers referenced. I asked for additional books and read those. I also checked out books from my public library. My best efforts fell short, and I never felt confident in my mastery of the topic. In this respect, my REUT experience was most frustrating. Nevertheless, for my part, the summer was most satisfiying, and I might attribute my frustrations to occasional miscommunication or lack of communication on my part. At some point, the gap between what I thought I needed to know and what I did know coupled with my inability to bridge that gap through self-study overwhelmed me to an extent that I felt that I was getting in the way of progress. In the end, to me the project director seemed spread to thin to address my growing list of questions, and I never quite developed a satisfactory and effective working relationship with any of the graduate assistants.

Though a research result eluded me, my main objective for the summer was achieved. My primary goal for the summer was to become a better high school teacher by becoming aware of my own weaknesses and becoming more aware of what the future holds for the high school student that is interested in pursuing post-secondary studies in math and science. Towards the former, I've realized a need to sharpen my skills and knowledge in computer programming, linear algebra, probability and statistics. These outcomes a expected going into the project. My focus during the past seven years has been teaching, coaching, and, hopefully playing a role in the process that shapes boys into young gentleman. In the future, perhaps, I hope to shift some of my focus towards representing the larger mathematics community through conference attendance, presentation, and personal study. As to the latter, which refers to how my experience will more directly benefit my students, I've taken at least a few lessons. I experienced the importance of preparing students to work in groups towards project goals, the value of having a foundation in computer programming, and the degree to which students will be expected to be conversant in the language of mathematics, whether it be spoken or written. In addition to the discipline-specific lessons that I learned this summer, I was reminded what it is like navigating an unfamiliar landscape. I had never heard of LaTeX before this summer, had virtually no background in biology prior to the start of the project, little experience with computer programming, and only basic linear programming. Since results depended on intermediate to advanced skills in all of the above, I was very much a student with a steep learning curve, feeling not unlike many of the students in my math classes must feel from time to time. Over the summer I kept notes about how I coped with the experience and feel certain that I will be more able to meet the needs of students that are feeling particularly

challenged in years to come.

References

[1] Peter Salamon Chrystian V. Irazoque and Annalinda Arroyo. Creating a phylogenetic tree using google ranks. Undergraduate and graduate posters in mathematics presented at the 2006 SACNAS National Conference.

[2] Peter Clote and Rolf Backofen. *Computational Molecular Biology*. Wiley, 2000.

[3] D.J. Lipman J. Ostell D.A. Benson, I Karsch-Mizrachi and D.L. Wheeler. Genbank. *Nucleic Acid Research*, 34 (Database issue):D16–20, 2006.

[4] Peter J. Waddell David L. Swofford, Gary J. Olsen and David M. Hillis. *Molecular Systematics*, chapter Inferring Phylogenies. Sinauer, Sunderland, 1996.

[5] William R. Taylor David T. Jones and Janet M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*, 8:275–282, 1992.

[6] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer, 2004.

[7] Bjarne Andresen Joseph M. Mahaffy Anca M. Segall David Mead Farooq Azam Forest Rohwer Mya Breitbart, Peter Salamon. Genomic analysis of an uncultured marine viral community. page 14250, 2002.

[8] Forest Rowher and Rob Edwards. The phage proteomic tree: A genome-based taxonomy for phage. *Journal of Bacteriology*, 184(16):4529–4535, 2002.

[9] Simon Whelan and Nick Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biologogy and Evolution*, 18(5):691–699, 2001.