

Fast Categorization of Bacteriophage Protein Families using Computer Graphics

By Christopher Kawasaki

August 10, 2005

B-SURE Summer Program 2005

Fast Categorization of Bacteriophage protein families using Computer Graphics

Abstract

Although there are many bacteriophages, only a small number of the ones existing in our soil have been characterized. In order to be able to use them for phage therapy, we need fast and accurate methods of categorizing them. The Hardies Laboratory chose to compare them with each other Secondary Structure Prediction, using SAM and Psipred. In order to accelerate the process of displaying the data in a visual format, we developed programs for taking SAM's and Psipred's output and displaying the aligned secondary structures onto a graphical image, which is made by Gbrowse. The result, is that hypotheses and experiments on secondary structure prediction programs using the same formatting standards have a much faster time to graphical image generation. With that, testing various aspects like accuracy of said secondary structure prediction programs becomes easier, as well as invaluable assistance in solving the hypothesis 'These phages are(are not) related.'

Introduction

Although there are many bacteriophages, only a small number of the ones existing in our soil have been characterized. A increasing field of study on phages is for use in phage therapy, or using phages to rid the human body of certain kinds of bacteria. Anything used on the human body to cure bacteriological infections would be a great bonus, however, it requires a great deal of knowledge of the phages in order to create or use such a form of treatment on a human. One way of determining functions of newly isolated phages is to compare them with already known phages. One method of comparing them for similar function is to compare the sequences of the proteins that the phages are made of. For example, the terminase, which is the protein whose function is to inject the DNA into a bacteria cell. This report will focus on the terminase family and subfamilies.

One computer program to compare phage proteins to each other is SAM, which is short for Sequence Alignment and Modeling System. SAM uses a complex set of algorithms in order to match the protein sequences given to it towards a common ancestor. SAM makes adjustments based on where there might be places where amino acid residues were lost or gained due to genetic mutation, and either puts in dashes to show where residues were lost, or lowercase letters where residues were probably added. Because SAM happens to be structure sensitive, and has large capabilities, it's more accurate than other alignment programs. It's one great drawback, however, is that it's too complex, so it needs a supercomputer to run in a reasonable amount of time. The availability of the supercomputer in the Bioinformatics Center allows the use of SAM in our investigations. When SAM is finished, it produces an a2m file, or Alignment to Model file. Because SAM can align something that is only chance matches, there is a need to confirm it's results.

In order to help confirm the results of SAM, the method of secondary structure prediction was used. Secondary Structure Prediction, which is predicting the spatial arrangement of amino acid residues that are near one another in the linear sequences, is one method used to determine the relationship between various proteins. The two most referred types of secondary structures are Alpha Helixes and Beta Strands. In secondary structure prediction, a section of the phage with a high density of amino acids that favor one type of secondary structure is predicted to be that particular secondary structure(Kihara, 2005). Modern techniques to predict secondary structures include machine learning techniques like neural networks, hidden Markov models, and support vector machines, which use data and information from known examples of secondary structures to program themselves(Kihara, 2005). Psipred, the advanced prediction program used in this work, uses a simplified neural network, yet provides a very accurate degree of prediction accuracy(Jones, 1999). When Psipred is finished executing, it produces an secondary structure 2 file(ss2 file).

One method a scientist trying to compare certain proteins for secondary structure similarities might use, takes both the ss2 files and a2m files and constructs a graph containing the aligned structures, which allows for the comparison of different proteins. While the process of hand drawing could take a couple of weeks, as well as several skilled man hours. Even by manually entering the data in another format to a computer graphics program might produce a bit neater of a graph, but it would still be time consuming. Because the project, as a component, needs to create hundreds or thousands of graphs. Because a pattern could be implemented to decide how the graphs was formed, therefore algorithms for computer programs (which will be named calc and mcalc respectfully) could be made in order to speed the process up considerably.

Results and Discussion

The result of running the programs developed is the graph in figure 2. It displays in less than thirty seconds, versus the two weeks for it to be done by hand and powerpoint, as seen in figure 4, which was produced previously by Dr. Hardies and was used to as a form of test data for how the graph should look if properly computed. In addition to allowing the faster generation of graphs, this method also allows for a fast and efficient method for comparing predicted secondary structures, as well as testing various subjects of the psipred and SAM programs itself. As can be seen by careful examination of the figure, the alpha helixes and beta coils line up for the most part, showing how they are related, even though some of them are really far from identical amino acid sequences.

To draw figures 2 & 3, as well as other graphs, we chose to use Gbrowse, short for Generic Genome Browser(Stein, 1599). Gbrowse is a free, open-source, image viewer created for mapping genes. We had to adapt the viewer to display more than one kind of secondary structure on a single track, or row, each containing a single protein on it, as well as program glyphs(a routine that draws shapes). For insertions, which is where we have extra sequences compared to the graph used, we needed a triangle whose point fell where the extra information was removed. For alpha helixes and beta strands, we used a helix glyph, which was a precise rectangle, because the already programmed glyph continued over the continued where they were supposed to end. The helix glyph was red for alpha helixes and blue for beta strands. For coils, we also used the helix glyph, but we made the dimensions of the rectangle really thin to produce a line. For the deletions, where we have fewer sequences compared to the alignment, we simply had a blank space free of other glyphs to indicate that. Now Gbrowse was ready to read a gff file, a file containing the coordinates of the glyphs. So the problem is now to convert the data into that format.

To convert the data from the SS2 and A2M file formats into the GFF file format, we wrote two programs, Calc and Mcalc. Both were created in perl, a computer programming language, for ease in both creation and adaptation into a perl cgi script format.

Calc, or Calculate, is designed to take three parameters, the ss2 filename, the a2m filename, and the GI number of the phage to be processed. It converts the data from the ss2 File using a finite state machine, into the protein's natural, or unaligned, coordinate system. Then it extracts the same sequence using the GI number from the a2m file, using a second finite state machine to locate the position and size of gaps and insertions where the alignment is to be altered. After finishing with both the two files, it adjusts the protein's natural coordinates by the alterizations from the a2m file accordingly, putting in the insertions and deletions from the a2m file into the coordinates extracted from the ss2 file. It then outputs the positions of all secondary structure positions, gaps, and inserts, in the same coordinate system used in gff files, which are piped into mcalc.

Mcalc, or Multiple Calculate, supervises calc and writes the final gff file. It is designed to be placed into the directory with all the ss2 files, an a2m file with all of the GI numbers, and a special file called family.txt, which has to be generated by the user. Family.txt contains the name of the a2m file, followed by each ss2 file, paired with it's GI number. First, mcalc reads family.txt, to get the

parameters for calc. Secondly, it executes calc, and pipes the results back into itself. Lastly, mcalc translates the coordinate system into a gff file, one track per ss2 file processed by calc. In addition, it adds titles for the dynamic creation of track(protein) groupings and maps into the Gbrowse configuration file. Now the gff file is ready to be fed into Gbrowse.

Gbrowse requires two things to produce the webpage; First, a gff file, and secondly, a configuration file, or conf file. The conf file contains all the header notes, such as the family title, bookmarks, website links, and Track labels. The track labels has been made dynamic by a special title glyph and some change made to mcalc. Currently, the maximum number of tracks set by the conf file used was 15 tracks.

This new software allows for a much faster development of progress in secondary structure prediction. As you can see in figure 2, The Terminase family has been divided into subgroups for comparison.

This software has the capability to not only generate graphs for comparison, but allows for faster performed scientific tests on both SAM and Psi-pred, as well as other secondary structure programs that use the same format. As an example analyzing the performance of Psipred, a second experiment was performed, analysing the triple beta strands in the motor domain 2(See figure 3). We divided the Sfi21 subfamily into several subgroupings of at least 5 proteins, and did secondary structure prediction on each of those groups. AS can be seen in figure 3, many of the subgroupings had a helix instead of a middle beta strand. The question is, does that mean those subfamilies evolved to have alpha helixes, or does it mean that psipred loses accuracy with small numbers of proteins. Those protein subgroups were combined and psipred was executed to perform secondary structure prediction over all of them. Figure 3, track rlt-com, has three beta strands, which indicates that the middle helix was mispredicted because of noise and lack of samples. It also supports the theory that as there is more samples in the data, the smaller the margin of error becomes.

In conclusion, calc and mcalc allow for fast and mechanical computational translation from standard data formats used by Psipred and SAM into a visual format, which can be used to test hypotheses and help provide figures for reports in far less time than performing these calculations by hand would provide. Improvements that could be made at a later date include fixing the graph to allow the different structures to link to a database in which they could give the user information. Also, programming mcalc to produce a dynamic configuration file for Gbrowse will be needed to reduce the amount of time spent on working on separate configuration files for different families of proteins.

Materials and Methods

Secondary Structure Prediction

The Hardies Lab used Psipred to determine the secondary structure prediction. Psi-pred uses psi-blast on a mirror of the NCBI NR protein database, which is updated nightly and maintained by the UTHSCSA Bioinformatics center. It then runs through a trained neural net to predict secondary structure. By comparing many similar phages, we then can use some advanced programs to use secondary structure prediction, Psipred being considered as one of the most reliable in the field(Volker,).

SAM

The SAM alignment of large terminase proteins were provided by Dr. Hardies. They were produced by iterative search of the NR database and alignment. Seed family was the Sfi21 family. The alignment was culled of fragmentary matches. Names of the viruses for each protein were added to the identifier. The tree describing the clustering of the sequences into subfamilies was provided by

Dr. Hardies. It was computed using the neighbor joining algorithm from the programming package named PAUP implemented at the UTHSCSA Bioinformatics Center.

Programs written

The programs developed are mcalc and calc, respectively. They are located on the bioinformatics center computer BCF, under /home/user/hardies/bin/Kawasaki. calc is called with the following usage: calc <ss2 file> <a2m file> <GI Number>, although it is usually called by mcalc rather than the user. Mcalc, when called, expects all of the ss2 and a2m files to be in the default, as well as a special instruction file called family.txt. Family.txt needs to contain the name of the a2m file on the first line; then, in order that they will appear on the display, the ss2filename with a space before the GI number unique to that sequence. Family.txt is created by the user. As seen in figure one, mcalc produces a gff file for use by an adapted version of Gbrowse, which is a open source Genome Browser (Stein, 1599).

To do subgroup analysis, the sequences of the a2m file were split from various subfamilies into subgroups based on their tree branches. The Location of the list of files used to determine their SAM alignment, as well as Gbrowse display links, SS2 files listed, and the tree used to map all of the families out can be found on Dr. Stephen Hardies Family Database, located at http://biochem.uthscsa.edu/hardies-bin/display_famv2.pl?db=schfamilies, under the table terminasev2.

Tree

In order to isolate subgroups of the tree and only align those, so that psipred will only use those sequences in it's prediction, a special blast formatted database corresponding to each subgroup had to be created. In order to do that, several steps were taken. Step one, was to take all of the names of a subgrouping from a list which was organized by the tree and put them into small lists of their own. Second, was to run the program sorttreea2m, which uses the following usage: sorttreea2m <main a2m file> <list file> <new a2m filename>, which filters a2m data into separate a2m files, based on those lists from step one. After we had the subgroup's a2m file, then we used the program formatdb from the NCBI Blast Toolkit to make it into a database format for psi-pred to read. It's usage statement is: formatdb -l <a2m> -pT -oT, and it creates several database subfiles used by psi-pred. The next thing done is to retrieve the sequence of an well recognized phage inside of this subgroup. To do that, we use the fastacmd to withdraw the information from the global database. By typing in -s <gi number> -d \$BLAST_DB/nr -o <fa filename>, the computer will retrieve the data needed from the NCBI global database and place it into a .fa file, which contains the sequence, by your name. The final step, is to run psi-pred, which is called by the command: runpsipred <fa file> <a2m file turned into a database>. This will generate an ss2 and Blast file by the same name as the fa file, only the extension will be changed. The ss2 file is now ready to be processed by mcalc and calc.

Acknowledgements

I'd like to thank Dr. Hardies, for all the help he's given me in the mainstay of working on this project, and for suppling me with programming manuals, and other various equipment.

Biography

Hughey R., Karplus K., and Krogh A. 2000. Sequence slignment and modeling software system. Technical Report UCSC-CRL-99-11, University of Californic. This Document available online at http://www.soe.ucsc.edu/research/compbio/papers/sam_doc/sam_doc.html

Hughey R, and A. Krogh A. 1996. Hidden Markov models for sequence analysis: Extension and analysis of the basic method, CABIOS, 12:95-107, 1996.

Jones, David T., 1999. Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *J Mol Biol.* 1999 Sep 17; 292(2):195-202.

Karplus K, Barrett C, Hughey R. 1998. Hidden Markov Models for detecting remote protein homologies. *Bioinformatics* 14:846-856. The UCSC SAM server is at <http://www.soe.ucsc.edu/research/compbio/sam.html>

Kihara, Daisuke, 2005. The effect of long-range interactions on the secondary structure formation of proteins. *Protein Science* (2005) Jun 29, 1-9.

Ross, B., Volker, E.A. EVA: large-scale analysis of secondary structure prediction. *Proteins*. 2001; Suppl 5:192-9.

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. The generic genome browser: a building block for a model organism system database. *Genome Res.* 2002 Oct;12(10):1599-1610.

Stryer, Lubert., 1995. *Biochemistry/Lubert Stryer.*— 4th ed; W. H. Freeman and Company, New York. pp 35.

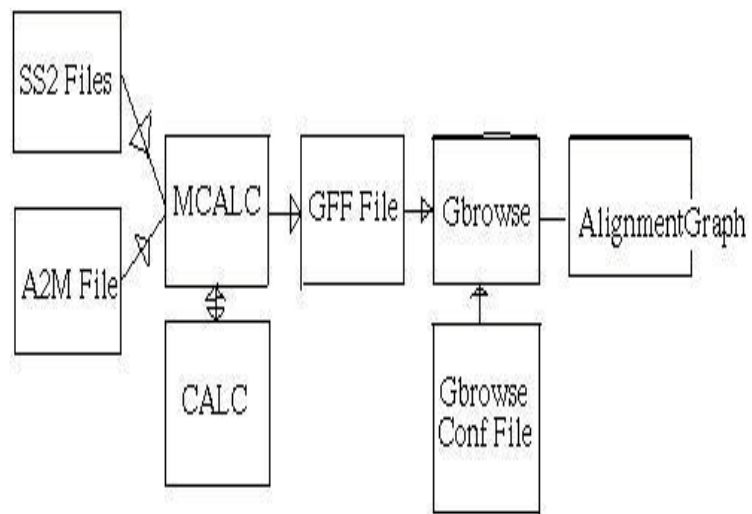


Fig.1 The proceeding steps from the standard format to visual representation on a graph.

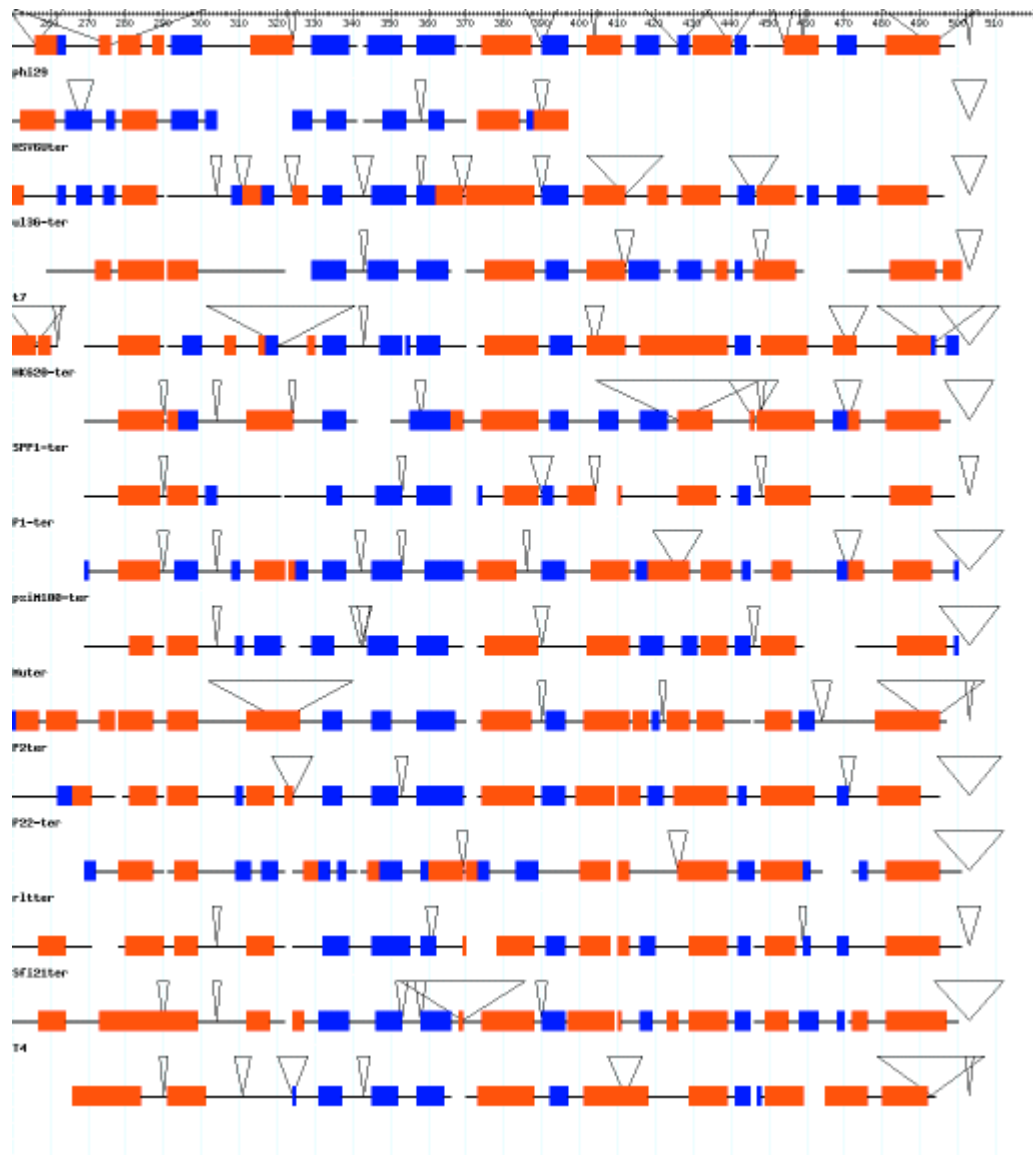


Fig. 2. Image containing the right half of the family Terminase and it's several subfamilies. Alpha Helixes are the red rectangles, and Beta strands are the blue figures. The line represents coil, and the triangle represent where data was inserted since it evolved from it's common ancestor, called insertions.

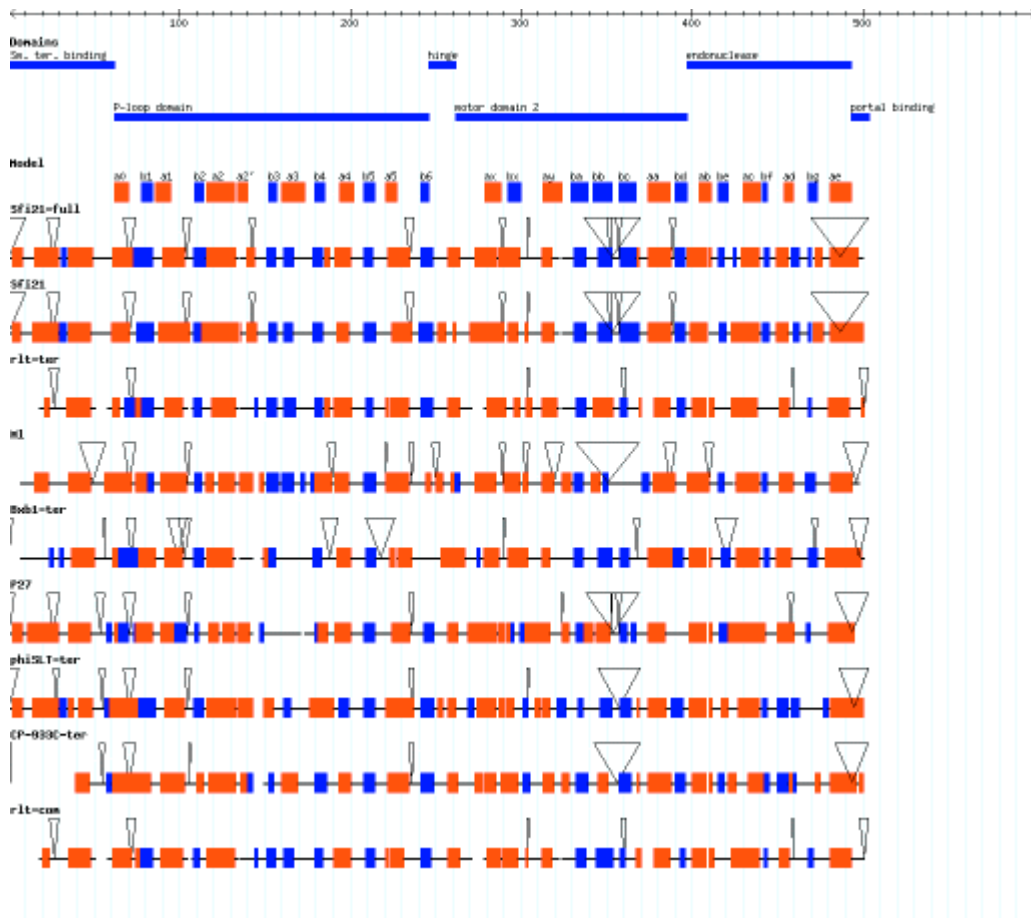


Fig. 3. An alignment graph showing Sfi21, a subfamily of terminase, and its various subgroupings. The last subgrouping, labeled by rit-com, is the experimental track.