中央研究院
資訊科學研究所
Institute of Information Science, Academia Sinica • Taipei, Taiwan, ROC

TR-IIS-06-007

Phylo-mLogo: An interactive multiple-logo visualization tool for large-number sequence alignments

Arthur Chun-Chieh Shih, D.T. Lee, Chin-Lin Peng, and Yu-Wei Wu

June 19, 2006 || Technical Report No. TR-IIS-06-007
http://www.iis.sinica.edu.tw/LIB/TechReport/tr2006/tr06.html

Phylo-mLogo: An interactive multiple-logo visualization tool for large-number sequence alignments

Arthur Chun-Chieh Shih[1], D.T. Lee[§1,2], Chin-Lin Peng[2] , and Yu-Wei Wu[1]

[1]Institute of Information Science, Academia Sinica, Taipei, 115, Taiwan
[2]Genomics Research Center, Academia Sinica, Taipei, 115, Taiwan

[§]Corresponding author
D.T. Lee
Distinguished Research Fellow & Director
Institute of Information Science, Academia Sinica
128 Academia Road, Section 2, Nankang, Taipei 115, Taiwan

Email: dtlee@ieee.org
Tel: +886-2-2788-3799 ext.2202
Fax: +886-2-2782-4814

Running title:
Phylo-mLogo: an alignment visualization tool

Keywords:
alignment visualization, sequence logo, influenza, avian influenza

Abstract

When aligning several hundreds or thousands of sequences, such as HIVs, dengue virus, and influenza viruses, to reconstruct the epidemiological history or to understand the mechanisms of epidemic virus evolution, how to analyze and visualize the large-number alignment results has become a new challenge for computational biologists. Although there are several tools available for visualization of very long sequence alignments, few of them are applicable to the large-number alignments. In this paper, we present a multiple-logo alignment visualization tool, called *Phylo-mLogo*, which allows the user to visualize the global profile of whole multiple sequence alignment and to hierarchically visualize homologous logos of each clade simultaneously. *Phylo-mLogo* calculates the variabilities and homogeneities of alignment sequences by base frequencies or entropies. Different from the traditional representations of sequence logos, *Phylo-mLogo* not only displays the global logo patterns of the whole alignment but also demonstrates their local logos for each clade. In addition, *Phylo-mLogo* also allows the user to focus only on the analysis of some important structurally or functionally constrained sites in the alignment selected by the user or by built-in automatic calculation. With *Phylo-mLogo*, the user can symbolically and hierarchically visualize hundreds of aligned sequences simultaneously and easily check the sites of their amino acid changes when analyzing large-number human or avian influenza virus sequences.

INTRODUCTION

Epidemic viruses, such as human immunodeficiency virus (HIV), influenza viruses, and dengue virus, continuously pose threats to human health, especially the recent outbreak of H5N1 avian influenza virus infection in human, which causes over 50% deaths among the 218 confirmed cases by WHO (http://www.who.int/csr/disease/avian_influenza/country/en/index.html) (Moya et al. 2004). Therefore, it is important and urgent for biologists to understand the mechanisms of epidemic virus evolution and the epidemiological history.

In addition to exponentially growing virus data available in GenBank, two ongoing large-scale sequencing projects of human and avian influenza viruses have released a number of complete virus genomes (Ghedin et al. 2005; Obenauer et al. 2006) and conducted several significant studies (Campitelli et al. 2006; Holmes et al. 2005; Obenauer et al. 2006). Different from those used for identification of conserved regions in comparative genomics, the sequences analyzed for epidemiology are usually much shorter and more conserved, and their number could be in the range of several hundreds to thousands. For examples, Homles *et al.* (2005) performed a phylogenetic analysis of 156 human H3N2 influenza A viruses and observed multiple co-circulating clades (Holmes et al. 2005). Recently, Campitelli *et al.* (2006) analyzed 685 human and avian sequences and found that viral genes appeared to be under strong purifying selection, with only the PB2, HA and NS1 genes under positive selection (Campitelli et al. 2006). Moreover, Obenauer *et al.* (2006) compared 4339 avian influenza virus genes and identified several novel clades never found before (Obenauer et al. 2006). Therefore, aligning large-number virus sequences can help researchers identify important polymorphic sites between different lineages and find out also the evolutionary histories and mutation trends of influenza viruses.

Sequence alignment and inference of the phylogenies is a standard procedure for analyzing virus sequences (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html). Based on the reconstructed phylogenetic relationship, the evolutionary histories of epidemic viruses can be inferred. Traditionally researchers are used to assigning numbers to all clades in the phylogenetic analysis of individual gene segments and using them to represent and compare genotypes across multiple viruses. However, when the number of analyzed sequences is in the hundreds, this approach cannot distinguish the differences between viral sequences from different strains (Obenauer et al. 2006). Moreover, the evolutionary changes of some specific sites, such as antigenic sites, cannot be directly observed by global phylogenetic analysis, short of checking the detailed alignment results. Therefore, how to provide efficient tools for biologists to analyze and visualize large-number sequence alignments of viruses has become a challenge for computational biologists.

In recent years there are several visualization tools of sequence alignments available in the public domain. Based on the visualization output, these tools can be divided into two categories: curve-based and sequence-logo-based. In the former category, tools, such as VISTA family (Shah et al. 2004), PipMaker (Schwartz et al. 2003; Schwartz et al. 2000), zPicture (Ovcharenko et al. 2004), and SinicView (Shih et al. 2006), are developed to either visualize individual alignment results or compare and evaluate assorted alignment results obtained by different tools. These tools are useful for visualizing very long sequence alignment results of a few sequences. However, for cases of large-number but short sequence alignments, they are impractical because some significant variations between sequences may be submerged by global scoring profiles which are calculated by either identical rates or sum-of-pair scores.

In the latter, sequence logos graphically represent the informative patterns of each individual site in a multiple sequence alignment. Thus, the sequence logos can assist users to discover and identify conserved patterns from multiple sequence alignments (Schneider and Stephens 1990). The original work was first proposed by Schneider and Stephens (1990). Ten years later, Crooks *et al*. (2004) performed an extension that incorporates additional features and options, called WebLogo (Crooks et al. 2004). For distinguishing the gaps and poorly conserved positions, LogoBar (Perez-Bercoff et al. 2006) was proposed to display protein sequence logos including not only amino acids but also gaps. These logo-based tools are very useful to globally visualize consensus patterns in a multiple sequence alignment. However, when the number of aligned sequences is in several hundreds or thousands, some significant local tendencies of mutations cannot be observed directly from these global logo-based profiles. In the analysis of influenza virus evolution, tracking the transitional changes of the amino acids at the epitope or receptor binding sites is very important because their changes could cause antigenic drift (Bush et al. 1999), affect viral transcription (Gabriel et al. 2005), and conduce mammalian adaptation (Subbarao et al. 1993). Furthermore, since their mutation rates are much faster than those of eukaryotes (Li 1997), observing dynamic evolutionary transitions of viruses can help the researchers examine the functional and evolutionary characteristics of influenza.

In this paper, we present a multiple-logo alignment visualization tool, called *Phylo-mLogo*, which allows the user to visualize the global profile of the whole multiple sequence alignment and to hierarchically visualize homologous logos of each clade simultaneously. *Phylo-mLogo* calculates the variabilities and homogeneities of aligned sequences by base frequencies or entropies. Different from the traditional representations of sequence logos, *Phylo-mLogo* not only displays the global logo patterns of the whole alignment but also demonstrates their local logos for each clade. In addition, *Phylo-mLogo* also allows the user to focus only on the analysis of some important structurally or functionally constrained sites in the alignment selected by the user or by built-in automatic calculation. With *Phylo-mLogo*, the user can symbolically and hierarchically visualize hundreds of aligned sequences simultaneously and easily check the sites of their amino acid changes when analyzing large-number sequences of human or avian influenza viruses.

RESULTS

In influenza viruses, the surface glycoproteins hemagglutinin (HA) is the most important target for the human immune system. Recent studies reveal that modifications of HA1, the immunogenic part of HA, accrue at a dramatic rate and also indicate that HA1 is undergoing diversifying or positive selection (Bush et al. 1999; Fitch et al. 1991; Plotkin and Dushoff 2003). Since their HA1 genes mutate so fast, the new variant strains of H3N2 tend to replace older ones quickly so as to cause annual outbreaks. Thus, to identify the sites under selection and their mutation trends in the HA genes is very important (Bush et al. 1999). In what follows, we will introduce two examples in the study of influenza HA genes to demonstrate how *Phylo-mLogo* can assist users to observe and analyze large-number sequences alignment results. The total numbers of alignment sequences in both of the examples are 453 and 207, respectively. The relationships of the aligned influenza sequences are acquired by human and avian in each example.

Example 1: 453 avian influenza HA genes

The spread of H5N1 avian influenza from China to Europe has raised global concern about their potential to infect humans and cause a pandemic. A more comprehensive collection of data and analysis of avian influenza sequences is critically needed for biologists and epidemiologists to find out the virulence and transmissibility of these viruses from avian species to humans. Thus, Obenauer *et al.* (2006) established the first large-scale sequencing effort to collect additional genomic data on the avian population of influenza A viruses (Obenauer et al. 2006). They introduced a proteotyping method to identify and number unique amino acid signatures, called *proteotypes*, for sequences that may or may not be distinguished by branches on a phylogenetic tree. They analyzed eight avian influenza genes and provided the proteograms to demonstrate the amino acid signatures within each clade (Figures S2-S9 in the supplementary material of (Obenauer et al. 2006)). Based on the observations, they concluded that the virus families tend to have multiple core conserved genes and that the surface glycoproteins, HA and NA, appear to be more freely exchanged than core proteins because of immune pressure (Obenauer et al. 2006).

In this part, we downloaded 437 avian influenza HA genes used for analysis in (Obenauer et al. 2006). To infer the phylogenetic tree of these sequences by MrBayes is very time consuming (Obenauer et al. 2006; Ronquist and Huelsenbeck 2003), we therefore observed the tree shown in Fig. S6 in (Obenauer et al. 2006) directly and constructed their phylogenetic relationship manually. The proteotypes of the analyzed sequences include p1.1, p2.1, p5.1-4, p6.1-6, p8.1, p9.1, p9.2, and p12.1. Based on these proteotypes, we first aligned the sequences of each proteotype and then aligned these proteotypes together, by ClustalW (Thompson et al. 1994). The total alignment length is 584.

Figure 2A shows the sequence logos and their phylogenetic tree, simultaneously. Different from other tools for tree visualization (http://www.genetics.wustl.edu/eddy/atv/), *Plylo-mLogo* displays the phylogenetic tree by using a standard file browser because this representation is more compact than that of the traditional tree visualization of the original phylogenetic tree as shown in Fig. 2B. Thus, the user can click on different clades shown in yellow background colors, like selecting different folders in a file browser, to visualize the sequence logos of the alignment at different levels.

Stevens *et al*. (2006) listed some conserved residues with the receptor binding domains of H1 and H5 serotypes that are implicated in receptor specificity, amino acid positions 183, 190, 193, 194, 216, 221, 222, and 225-8 (Stevens et al. 2006) of which the corresponding positions in our example are 205, 212, 215, 216, 238, 243, 244, and 247-250, respectively. Then, we compared these sequence logos between different proteotypes. To avoid confusion in this example, we used the original positions shown in Stevens *et al*. (2006) in the following discussion. As shown in Fig. 2C, the amino acids at residue sites 194, 225, and 228 are almost conserved across H1, H2, H5, H6, H8, H9, and H12 serotypes. If we only consider H1, H2, and H6, the same clades with H5 (Glaser et al. 2005; Obenauer et al. 2006), the amino acids at sites 183, 190, 194, 225, 226, and 228 are almost the same across these serotypes. Interestingly, we found that the majority at residue 221 of the proteotype p5.3 is the amino acid S not P in the reference avian strain, A/Duck/Singapore/Q-F119-3/1997 (H5 serotype) used for comparison in (Stevens et al. 2006), while 221S has been fixed in the human H5N1 strain, A/Vietnam/1203/2004. Since all sequences in p5.3 belong to avian H5N1, it appears that the 221S (polar amino acid) has almost taken over 221P (non-polar amino acid) in the avian H5N1 population. At residue 216, the polymorphisms of amino acids, K, E, and R, are all found in proteotype p5.3 while the amino acids are 216E and 216R for A/Duck/Singapore/Q-F119-3/1997 and A/Vietnam/1203/2004, respectively. Compared with those for H1, H2, and H6, 216K seems to be an advantageous mutation and may be probably fixed at the site in the avian H5N1 strains later. Since the amino acids K and R are positive charged but E is negative charged, the receptor specificity of residue 216 could definitely be changed in the avian H5N1 population consequently.

Moreover, the cleavability of the HA molecule for avian influenza A viruses plays a major role in virulence in birds and the amino acid sequence at the HA cleavage site, PQRERRRKKR/G, is considered as the most important pattern (Hatta et al. 2001). Between sites 352 and 355, we also identified this amino acid profile, PQRERRRKKR, in the sequence logo of p5.3 in which almost all 132 HA sequences belong to H5N1, while only few of H5N1 isolates are grouped in other proteotypes p5.1, p5.2, and p5.4. Thus, this pattern seems to appear only in H5N1.

Briefly, *Phylo-mLogo* can assist users to compare and visualize the changes of polymorphisms and indel events across different clades or subtypes of large-number sequence alignment so that users could speculate possible evolutionary and functional mechanisms to examine their hypotheses further.

Example 2: 207 human influenza H3N2 isolates collected from New York (1998-2004)

The influenza H3N2 virus has been infecting the human population over a long timescale as a major cause of morbidity and morality (Horimoto and Kawaoka 2005). To comprehensively study the virus evolution, Ghedin *et al*. (2005) sequenced 207 complete genomes of human H3N2 influenza A viruses collected between 1998 and 2004 from New York State, USA. These large-scale data could provide a comprehensive look at an influenza virus population across several seasons within a constrained geographical area (Ghedin et al. 2005). Different from using the phylogeny inference of partial genomes in their prior study (Holmes et al. 2005), the ten main proteins in these 207 genomes were organized according to their sampling seasons and demonstrated all amino acids of those sites having changes in at least three isolates in a colorful table-like map as shown in Fig. 1 in (Ghedin et al. 2005) or its Supplementary Fig. 1. They found a number of important mutations in the data that may affect receptor-binding affinity and potentially increase viral replication efficiency (Ghedin et al. 2005).

In this example, we downloaded the HA genes of 207 influenza virus genomes from NCBI (http://www.ncbi.nlm.nih.gov/) based on the GenBank accession numbers listed in Ghedin *et al.*'s supplementary data (Ghedin et al. 2005). Considering the sampling influenza seasons, we created a relationship tree with five seasonal groups, Nov. 1998- Mar.1999, Oct.1999-Mar. 2000, Nov. 2001-Mar. 2002, Feb.-May. 2003, and Oct. 2003-Feb. 2004, as an input tree and then aligned the amino acid sequences belonging to the same group. Finally, all aligned sequence profiles were aligned together for the cross-season comparison. After eliminating the first 16 sites of gap-rich regions, the total length of the alignment is 550.

Since most sites in the sequences across seasons are highly conserved, we may only need to observe those changing sites. Using the built-in automatic site selection provided in *Phylo-mLogo*, Fig. 3A shows 52 sites of the alignment result on the Root Logo View (the first row) and the Clade Logos View (the 2nd to 6th rows) when the cutting threshold is set to 0.99, i.e. the maximum frequencies of all amino acids at the sites are less than 0.99. If the threshold is set to 0.86, the total columns of the logo sites can be reduced to only 24 (Fig. 3B). Thus, the user can examine the heterogeneities of the whole alignment, 113,850 bases in total, in a single screen page.

In Fig. 3B, the user can easily see that the residues 25, 50, 45, 83, 131, 155, 156, 202, 222, 225, 386, and 530 simultaneously drifted their amino acids in some isolates during the season 2003 (only from March 2003 to May 2004) and all of these mutations soon were fixed in the whole population in the 2003-2004 season. These residues are important either for antibodies, such as 155 and 156, or for efficient replication, such as 131, 222, 225, and 226 (Jin et al. 2005). Holmes *et al.* (2005) have described that an epidemiologically significant reassortment appeared in this short season and caused the annual vaccine for that year to have limited effectiveness

(Ghedin et al. 2005; Holmes et al. 2005).

In addition, the residues 5V, 33H, 92T, and 271N only appeared in the season 1999-2000 while the residues 106V, 347M, and 529I suddenly appeared in the season 2001-2002 but were respectively replaced by the original residues 106A, 347V, and 529V later after 2003. It may imply that the original residues 5G, 33Q, 92K, 106A, 271D, 347V, and 529V on these sites are more advantageous than others. Furthermore, the residue 144, located at the epitope A region, has three substitutions from 1998 to 2003. Because the site was mutated so frequently, it implies that this site is under strong immune selection. In the 2003-2004 season, the mutations 126D and 547D have emerged and seem to replace 126N and 547N in the population, respectively. It may therefore be worth paying attention to whether these mutations will be fixed later and then cause the annual vaccines to become ineffective.

DISCUSSION

In population genetics, the selective sweep is the process by which a new advantageous mutation eliminates or reduces variation in linked neutral sites, as it increases in frequency in the population (Nielsen 2005). As circulated influenza H3N2 HA genes shown in Example 2, we can observe several sites changing under this process during flu season. When the mutations take the natural advantage, these variations will rapidly emerge and be fixed for a few years because the viruses can survive in human hosts. These mutations usually locate at important functional sites at which amino acid changes probably affect receptor-binding affinity and potentially increase viral replication efficiency for improving their fitness in the human host (Ghedin et al. 2005). Moreover, the identification of these mutational trends can help epidemiologists examine whether the recommended vaccine strains currently in worldwide circulation will continue to dominate in the next epidemic season in maintaining its effectiveness.

To date, deluged by the rate and complexity of completed genomic sequences, the need to align not only longer sequences but also larger number of sequences becomes more urgent. When comparing large-scale sequences, one of the objectives is to identify common conserved regions across species. Usually, the number of sequences is small but the total number of bases could be up to tens of millions (Thomas et al. 2003). In contrast, when aligning large-number sequences we focus on finding the heterogeneities of sequences that belong to the same species or population. Thus, the visualization target for large-scale sequence alignments is usually to demonstrate homologous regions and identify their conservations while that for large-number sequence alignments is to focus on the heterogeneities of the whole alignment and to exhibit common properties of these heterogeneous parts. Thus, *Phylo-mLogo* is suitable for users when they want to observe the heterogeneities of a large-number sequence alignment as well as to look at the homogeneities of these heterogeneous sites.

Generally speaking, *Phylo-mLogo* provides an effective and efficient representation for hierarchical visualization of large-number influenza virus sequences. The proposed representation can help users observe directly the heterogeneities and mutation trends from hundreds of sequences. However, the sequence-logo-based approach is site independent and does not indicate the relationship between strains. Therefore, combining *Phylo-mLogo* with graphically detailed maps like proteotype (Obenauer et al. 2006) and colourful amino acid map in (Ghedin et al. 2005) could assist users to understand the correlative relationship between different strains.

The maximum number of sequences and total length of alignment allowed in *Phylo-mLogo* are dependent on the capacity of internal memory and Java setting of the computer used. In practice, if *Phylo-mLogo* performs on a 3GHz Pentium4 PC with 1GB RAM, it is suggested that the maximum length of alignment be less than 500 bases and the maximum number of sequences smaller than 500 to obtain a good performance on both sequence logo calculation and displaying speed.

Deluged by the increasing number of virus sequences, how to visualize large-number sequence alignment results has become a new challenge for computational biologists. In this paper, we have presented a multiple-logo alignment visualization tool. Using *Phylo-mLogo*, the user can visualize the global profile of the whole multiple sequence alignment and to hierarchically examine homologous logos of each clade simultaneously. With *Phylo-mLogo*, the user can symbolically and hierarchically visualize hundreds of aligned sequences simultaneously and easily check the potential sites under different selection pressures, as demonstrated in the analysis of large-number human or avian influenza virus sequences. More information of *Phylo-mLogo* can be found at URL http://biocomp.iis.sinica.edu.tw/ the website of the Computational Genomics Lab., Genomics Research Center and Institute of Information Science, Academia Sinica.

METHODS

There are a viewing section and a panel control section in *Phylo-mLogo*. The viewing section includes Root Logo View, Clade Logos View, and Detailed Text Alignment View, while the control section provides the user with a suite of versatile control functions for visualizing the alignment results from different perspectives. In what follows, we will introduce the characteristics and functionality of *Phylo-mLogo* in more detail.

Scoring Methods in *Phylo-mLogo*

In *Phylo-mLogo*, the scoring methods to calculate the variabilities and homogeneities of aligned sequences are based on entropy or base frequencies. Sequence logos graphically represent the informative patterns in a multiple sequence alignment (Crooks et al. 2004; Schneider and Stephens 1990). Schneider and Stephens (1990) defined the *sequence conservation* at a fixed position in the alignment as the difference between the maximum possible entropy and the entropy of the observed symbol distribution (Schneider and Stephens 1990):

$$R_{seq} = S_{\max} - S_{obs} = \log_2 N - (-\sum_{i=1}^{N} p_i \log_2 p_i) ,$$

where p_i is the observed frequency of symbol i at a fixed position and N is the number of distinct symbols for the given sequences types (Schneider and Stephens 1990). The maximum sequence conservation per site is $\log_2 4=2$ bits for nucleotide sequences and $\log_2 20 \sim 4.32$ bits for amino acid sequences (Crooks et al. 2004). A sequence logo shows each base by the total number of bits of information multiplied by the relative occurrence of the nucleotide or amino acid at the position. Sequence logos enable fast and intuitive visual assessment of pattern characteristics applied to many fields (Wasserman and Sandelin 2004). However, in some not highly conserved cases, the entropies of the sequences are usually lower or just close to half of maximum sequence conservation so that the generated sequence logos may be compressed together and not conspicuous. Thus, the magnitudes of the sequence logos in some literature are not multiplied by their entropies and only the relative occurrence frequencies of the bases are shown at the position (Smith et al. 2005a; Smith et al. 2005b). For convenience of making a distinction, we call those magnitudes without multiplied by the entropies the *frequency sequence logos* and those multiplied by the entropies the *entropy sequence logos*. The user can select different type of sequence logos shown in the Root Logo View and Clade Logos View sections.

Viewing Section in *Phylo-mLogo*

As shown in Fig.1, the sequence logo in the Root Logo View represents the base information of the whole alignment and the heights of each logo are the occurrence frequencies. If the user selects the entropy option, all heights of the logos will be multiplied by the relative entropies. Below the logos in this View, the positions in the whole alignment are also marked so as to assist users in comparisons with others in the literature.

In the Clade Logos View, the panels show different sequence logos of the aligned sequences belonging to selected clades. By observing the graphical results, it is much more intuitive and straightforward to judge and identify the selective or evolutionary trends of alignment results between different phylogeny, regionalization, or seasonality. When the user clicks the colored bar above the logo graphics, the Detailed Text Alignment View will automatically show the detailed alignment result in a colored text format where identical characters are shown.

Panel Control Section in *Phylo-mLogo*

In the control section, there are three panels for selecting the sequence logos from root to clades and the sites from totality to partiality: Phylogeny/Relationship Viewing Control, Site Selection Control, and Viewing Range Control. Phylogeny/Relationship Viewing Control shows the relationship structure of input aligned sequences. The structure could be the phylogenetic tree for cross-species or different lineage sequences, the regionalization for polymorphic sequences, or the seasonalities for epidemic influenza virus sequences. Because showing a scaled tree usually takes up a lot of space on a screen, in *Phylo-mLogo,* Phylogeny/Relationship Viewing Control demonstrates these relationships by a hierarchical file system browser, an unscaled tree, in which each node (a clade) represents a group of well-aligned sequences and the related *child* nodes will be expanded once the user clicks the *parent* node. As a file or directory name shown in the file browser, there is a name following each node given by the user or by built-in automatic assignment. When the user clicks the name part of either a parent or child node, the sequence logo associated with this node will be calculated and the result shown in the Clade Logos View. When the user clicks the name part again (a toggle switch), its sequence logo will become invisible.

In some cases, the user may be interested in only some specific regions. *Phylo-mLogo* also provides a function, called Viewing Range Control, to let users select the range of interest and demonstrate only the local sequence logo in the logo viewing panels. Moreover, *Phylo-mLogo* also allows users to select some important structurally and functionally constrained or informative sites in the aligned sequences and then visualize only the sequence logos of these specific sites. In Site Selection Control, these sites can be selected by the user manually or by built-in automatic calculation which is based on selection of either high frequency bases in low conservation cases or low frequency ones in high conservation cases. Thus, when the length of alignment result is long or the sequences are highly similar or only partially conserved, the user can observe only those informative sites rather than examine the whole alignment.

Other characteristics of *Phylo-mLogo*

In *Phylo-mLogo*, all graphics of the sequence logos can be exported to portable color image files. Furthermore, *Phylo-mLogo* is implemented entirely in Java language to ensure portability across major platforms. The execution procedure of the standalone *Phylo-mLogo* is quite straightforward. When launched, the user will be prompted two options to read user’s phylogenetic or relationship tree files and alignment results from the local disk. For more details of the installation and operations please visit our

Website (http://biocomp.iis.sinica.edu.tw).

ACKNOWLEDGMENTS

We thank Dr. H.-K. Tsai for valuable discussions and Messrs. Tze-Chang Shiao and Yi-Chiao Fang for assisting in the collection and alignment of the influenza virus sequences. This work was supported by the Center for Disease Control, Taiwan, under the grant No. DOH95-DC-1401, Thematic Project of Academia Sinica, Taiwan, the National Science Council of Taiwan under the grants No. NSC 94-2213-E-001-029 and NSC-94-2752-E-002-005-PAE, and by the Institute of Information Science, and the Genomics Research Center of Academia Sinica in Taiwan.

FIGURE LEGENDS

Figure 1

The sequence logos of the influenza virus HA isolates from New York for the whole periods and different influenza seasons, Nov. 1998- Mar.1999, Oct.1999-Mar. 2000, Nov. 2001-Mar. 2002, Feb.-May. 2003, and Oct. 2003-Feb. 2004. In the Viewing Section, the sequence logo shown in Root Logo View represents the frequencies of the amino acids over the whole aligned sequences while those shown in Clade Logos Views demonstrate the logos for the sequences belonging to the same clade.

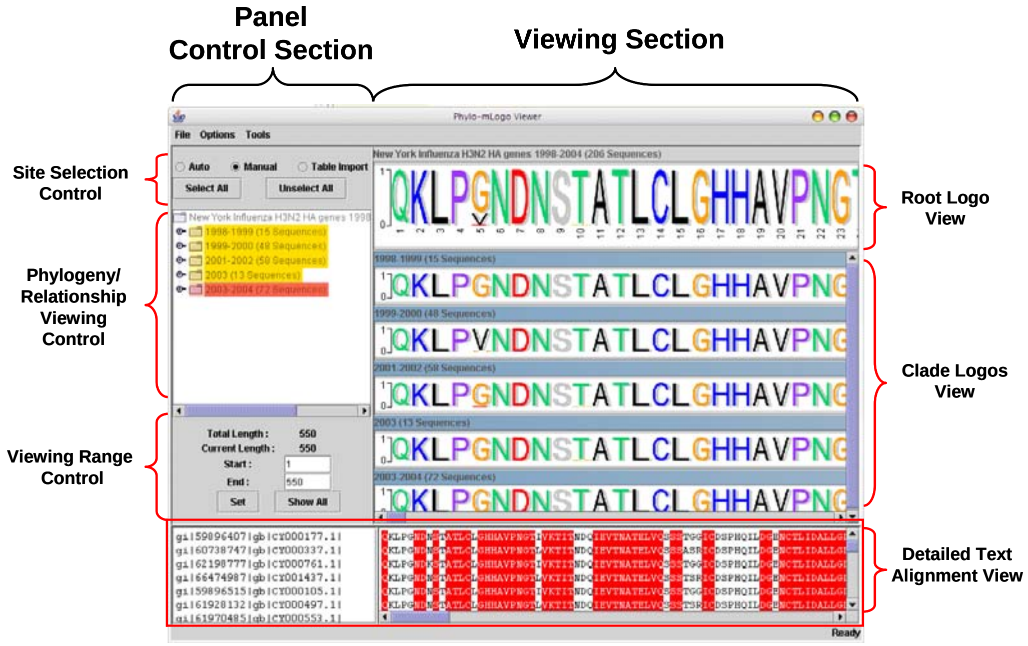
Figure 2

The sequence logos of the root and each clade of 453 avian influenza HA isolates. (a) The screenshot of *Phylo-mLogo* result. (b) The original phylogenetic tree. (c) The important sites located at the receptor binding domain and mapped to the alignment data.

Figure 3

The graphic outputs of the selected sequence logos of the HA genes in 206 New York H3N2 isolates by different cutting thresholds: A. 0.99; B. 0.86, where the first row shows the Root Logo View and the 2nd to 6th rows show the Clade Logos View.

Figure 1

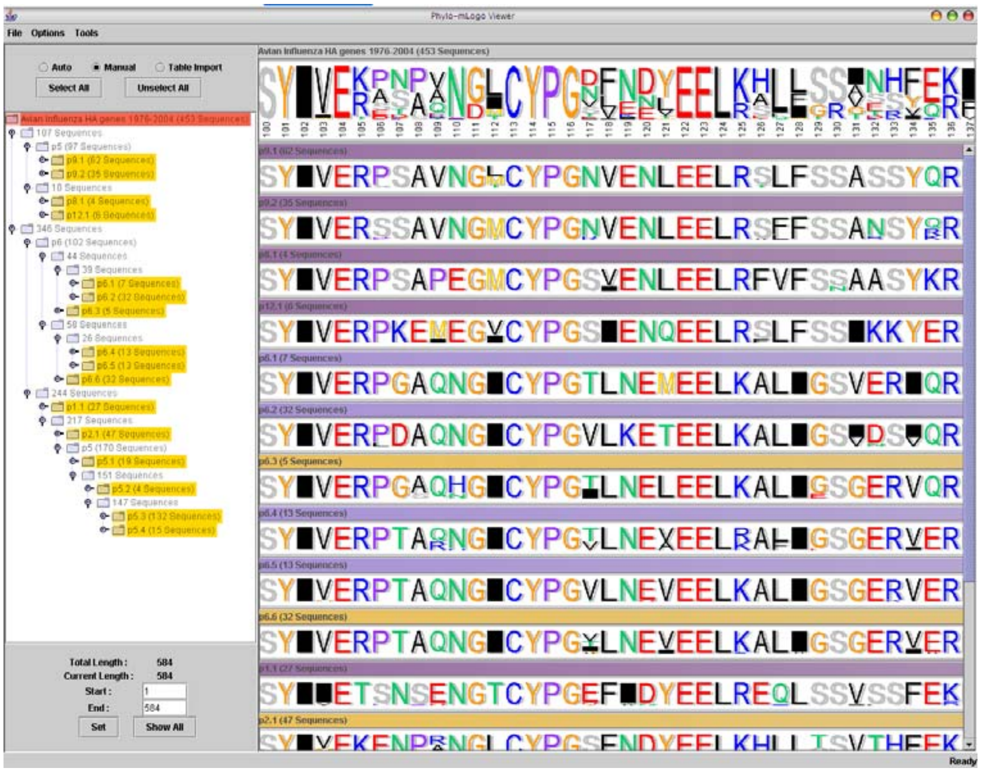


Total Length : 550
Current Length : 550
Start : 1
End : 550

```
gi|59896407|gb|CY000177.1|  QKLPGNDNSTATLCLGHHAVPNGTIVKTITNDQIEVTNATELVQSSSTGGICDSPHQILDGENCTLIDALLG
gi|60738747|gb|CY000337.1|  QKLPGNDNSTATLCLGHHAVPNGTLVKTITNDQIEVTNATELVQSSSASRICDSPHQILDGENCTLIDALLG
gi|62198777|gb|CY000761.1|  QKLPGNDKSTATLCLGHHAVPNGTIVKTITNDQIEVTNATELVQSSSTGGICDSPHQILDGENCTLIDALLG
gi|66474987|gb|CY001437.1|  QKLPGNDNSTATLCLGHHAVPNGTLVKTITNDQIEVTNATELVQSSSTSRICDSPHQILDGENCTLIDALLG
gi|59896515|gb|CY000105.1|  QKLPGNDNSTATLCLGHHAVPNGTIVKTITNDQIEVTNATELVQSSSTGGICDSPHQILDGENCTLIDALLG
gi|61928132|gb|CY000497.1|  QKLPGNDNSTATLCLGHHAVPNGTLVKTITNDQIEVTNATELVQSSSTSRICDSPHQILDGENCTLIDALLG
gi|61970485|gb|CY000553.1|
```

Figure 2

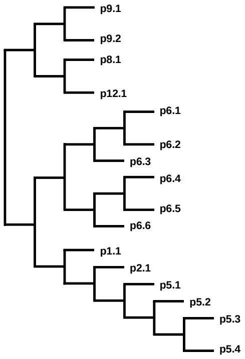
A



- Avian Influenza HA genes 1976-2004 (453 Sequences)
  - 107 Sequences
    - p5 (97 Sequences)
      - p9.1 (62 Sequences)
      - p9.2 (35 Sequences)
    - 10 Sequences
      - p8.1 (4 Sequences)
      - p12.1 (6 Sequences)
  - 346 Sequences
    - p6 (102 Sequences)
      - 44 Sequences
        - 39 Sequences
          - p6.1 (7 Sequences)
          - p6.2 (32 Sequences)
        - p6.3 (5 Sequences)
      - 58 Sequences
        - 26 Sequences
          - p6.4 (13 Sequences)
          - p6.5 (13 Sequences)
        - p6.6 (32 Sequences)
    - 244 Sequences
      - p1.1 (27 Sequences)
      - 217 Sequences
        - p2.1 (47 Sequences)
        - p5 (170 Sequences)
          - p5.1 (19 Sequences)
          - 151 Sequences
            - p5.2 (4 Sequences)
            - 147 Sequences
              - p5.3 (132 Sequences)
              - p5.4 (15 Sequences)

Total Length : 584
Current Length : 584
Start : 1
End : 584

B



C

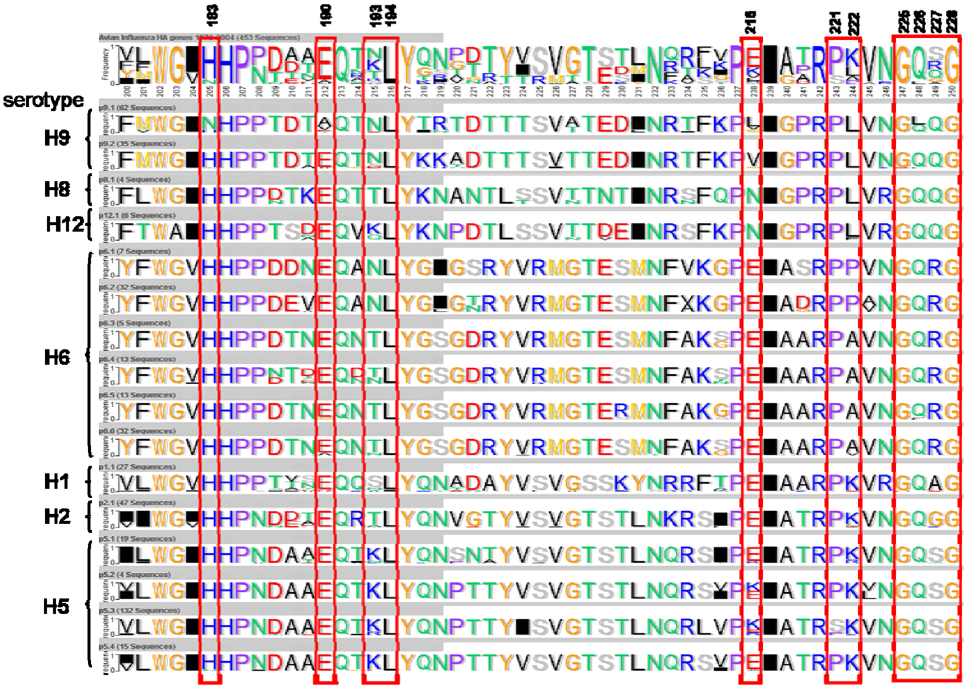
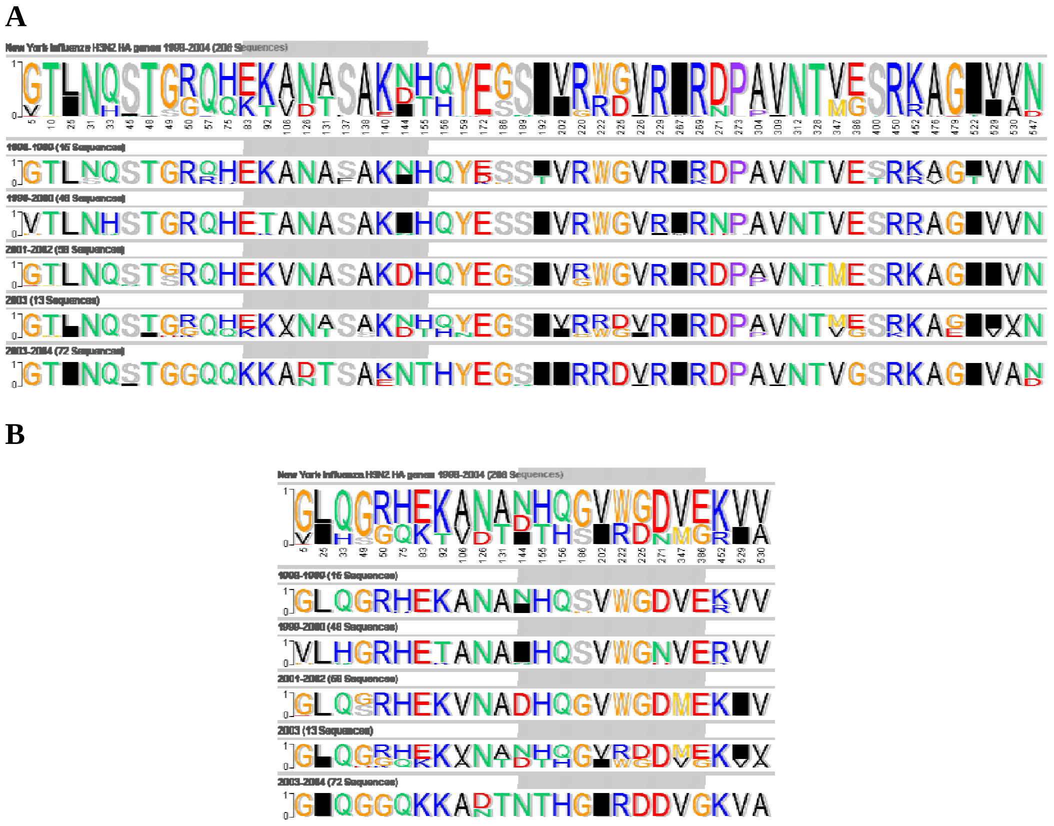
183
190
193
194
216
221
222
225
226
227
228
Avian Influenza HA genes 1976-2004 (453 Sequences)
Frequency
serotype
H9
p9.1 (62 Sequences)
p9.2 (35 Sequences)
H8
p8.1 (4 Sequences)
H12
p12.1 (6 Sequences)
H6
p6.1 (7 Sequences)
p6.2 (32 Sequences)
p6.3 (5 Sequences)
p6.4 (13 Sequences)
p6.5 (13 Sequences)
p6.6 (32 Sequences)
H1
p1.1 (27 Sequences)
H2
p2.1 (47 Sequences)
H5
p5.1 (19 Sequences)
p5.2 (4 Sequences)
p5.3 (132 Sequences)
p5.4 (15 Sequences)

Figure 3



REFERENCES

Bush, R.M., C.A. Bender, K. Subbarao, N.J. Cox, and W.M. Fitch. 1999. Predicting the evolution of human influenza A. *Science* **286:** 1921-1925.

Campitelli, L., M. Ciccozzi, M. Salemi, F. Taglia, S. Boros, I. Donatelli, and G. Rezza. 2006. H5N1 influenza virus evolution: a comparison of different epidemics in birds and humans (1997-2004). *J Gen Virol* **87:** 955-960.

Crooks, G.E., G. Hon, J.M. Chandonia, and S.E. Brenner. 2004. WebLogo: a sequence logo generator. *Genome Res* **14:** 1188-1190.

Fitch, W.M., J.M. Leiter, X.Q. Li, and P. Palese. 1991. Positive Darwinian evolution in human influenza A viruses. *Proc Natl Acad Sci U S A* **88:** 4270-4274.

Gabriel, G., B. Dauber, T. Wolff, O. Planz, H.D. Klenk, and J. Stech. 2005. The viral polymerase mediates adaptation of an avian influenza virus to a mammalian host. *Proc Natl Acad Sci U S A* **102:** 18590-18595.

Ghedin, E., N.A. Sengamalay, M. Shumway, J. Zaborsky, T. Feldblyum, V. Subbu, D.J. Spiro, J. Sitz, H. Koo, P. Bolotov, D. Dernovoy, T. Tatusova, Y. Bao, K. St George, J. Taylor, D.J. Lipman, C.M. Fraser, J.K. Taubenberger, and S.L. Salzberg. 2005. Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature* **437:** 1162-1166.

Glaser, L., J. Stevens, D. Zamarin, I.A. Wilson, A. Garcia-Sastre, T.M. Tumpey, C.F. Basler, J.K. Taubenberger, and P. Palese. 2005. A single amino acid substitution in 1918 influenza virus hemagglutinin changes receptor binding specificity. *J Virol* **79:** 11533-11536.

Hatta, M., P. Gao, P. Halfmann, and Y. Kawaoka. 2001. Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. *Science* **293:** 1840-1842.

Holmes, E.C., E. Ghedin, N. Miller, J. Taylor, Y. Bao, K. St George, B.T. Grenfell, S.L. Salzberg, C.M. Fraser, D.J. Lipman, and J.K. Taubenberger. 2005. Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biol* **3:** e300.

Horimoto, T. and Y. Kawaoka. 2005. Influenza: lessons from past pandemics, warnings from current incidents. *Nat Rev Microbiol* **3:** 591-600.

Jin, H., H. Zhou, H. Liu, W. Chan, L. Adhikary, K. Mahmood, M.S. Lee, and G. Kemble. 2005. Two residues in the hemagglutinin of A/Fujian/411/02-like influenza viruses are responsible for antigenic drift from A/Panama/2007/99. *Virology* **336:** 113-119.

Li, W.-H. 1997. *Molecular Evolution*. Sinauer Press, Sunderland, MA.

Moya, A., E.C. Holmes, and F. Gonzalez-Candelas. 2004. The population genetics and evolutionary epidemiology of RNA viruses. *Nat Rev Microbiol* **2:** 279-288.

Nielsen, R. 2005. Molecular signatures of natural selection. *Annu Rev Genet* **39:** 197-218.

Obenauer, J.C., J. Denson, P.K. Mehta, X. Su, S. Mukatira, D.B. Finkelstein, X. Xu, J. Wang, J. Ma, Y. Fan, K.M. Rakestraw, R.G. Webster, E. Hoffmann, S. Krauss, J. Zheng, Z. Zhang, and C.W. Naeve. 2006. Large-scale sequence analysis of avian influenza isolates. *Science* **311:** 1576-1580.

Ovcharenko, I., G.G. Loots, R.C. Hardison, W. Miller, and L. Stubbs. 2004. zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Res* **14:** 472-477.

Perez-Bercoff, A., J. Koch, and T.R. Burglin. 2006. LogoBar: bar graph visualization of protein logos with gaps. *Bioinformatics* **22:** 112-114.

Plotkin, J.B. and J. Dushoff. 2003. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. *Proc Natl Acad Sci U S A* **100:** 7152-7157.

Ronquist, F. and J.P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572-1574.

Schneider, T.D. and R.M. Stephens. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18:** 6097-6100.

Schwartz, S., L. Elnitski, M. Li, M. Weirauch, C. Riemer, A. Smit, E.D. Green, R.C. Hardison, and W. Miller. 2003. MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res* **31:** 3518-3524.

Schwartz, S., Z. Zhang, K.A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. 2000. PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* **10:** 577-586.

Shah, N., O. Couronne, L.A. Pennacchio, M. Brudno, S. Batzoglou, E.W. Bethel, E.M. Rubin, B. Hamann, and I. Dubchak. 2004. Phylo-VISTA: interactive visualization of multiple DNA sequence alignments. *Bioinformatics* **20:** 636-643.

Shih, A.C., D.T. Lee, L. Lin, C.L. Peng, S.H. Chen, Y.W. Wu, C.Y. Wong, M.Y. Chou, T.C. Shiao, and M.F. Hsieh. 2006. SinicView: A visualization environment for comparisons of multiple nucleotide sequence alignment tools. *BMC Bioinformatics* **7:** 103.

Smith, A.D., P. Sumazin, D. Das, and M.Q. Zhang. 2005a. Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics* **21 Suppl 1:** i403-412.

Smith, A.D., P. Sumazin, and M.Q. Zhang. 2005b. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci U S A* **102:** 1560-1565.

Stevens, J., O. Blixt, T.M. Tumpey, J.K. Taubenberger, J.C. Paulson, and I.A. Wilson. 2006. Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. *Science* **312:** 404-410.

Subbarao, E.K., W. London, and B.R. Murphy. 1993. A single amino acid in the PB2 gene of influenza A virus is a determinant of host range. *J Virol* **67:** 1761-1764.

Thomas, J.W., J.W. Touchman, R.W. Blakesley, G.G. Bouffard, S.M. Beckstrom-Sternberg, E.H. Margulies, M. Blanchette, A.C. Siepel, P.J. Thomas, J.C. McDowell, B. Maskeri, N.F. Hansen, M.S. Schwartz, R.J. Weber, W.J. Kent, D. Karolchik, T.C. Bruen, R. Bevan, D.J. Cutler, S. Schwartz, L. Elnitski, J.R. Idol, A.B. Prasad, S.Q. Lee-Lin, V.V. Maduro, T.J. Summers, M.E. Portnoy, N.L. Dietrich, N. Akhter, K. Ayele, B. Benjamin, K. Cariaga, C.P. Brinkley, S.Y. Brooks, S. Granite, X. Guan, J. Gupta, P. Haghighi, S.L. Ho, M.C. Huang, E. Karlins, P.L. Laric, R. Legaspi, M.J. Lim, Q.L. Maduro, C.A. Masiello, S.D. Mastrian, J.C. McCloskey, R. Pearson, S. Stantripop, E.E. Tiongson, J.T. Tran, C. Tsurgeon, J.L. Vogt, M.A. Walker, K.D. Wetherby, L.S.

Wiggins, A.C. Young, L.H. Zhang, K. Osoegawa, B. Zhu, B. Zhao, C.L. Shu, P.J. De Jong, C.E. Lawrence, A.F. Smit, A. Chakravarti, D. Haussler, P. Green, W. Miller, and E.D. Green. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788-793.

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673-4680.

Wasserman, W.W. and A. Sandelin. 2004. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5:** 276-287.