# Interactive Poster: Effective Display of Conserved Domains on a Multiple Sequence Alignment

Andrew D. Lindeman*, Susan M. Bridges†, T.J. Jankun-Kelly‡
Department of Computer Science and Engineering and the Institute for Digital Biology
Mississippi State University, MS 39762

## Abstract

Multiple sequence alignment (MSA) is used to explore the similarity of several related protein sequences by providing a near optimal alignment of the characters in each sequence. Biologists require effective visualization of these alignments as part of their analysis. Although tools such as Jalview have been developed that provide a detailed view of different aspects of the alignments and metadata such as conserved domains, these tools do not automatically download the metadata via the web and the alignment cannot be viewed at varying levels of detail. We have developed a prototype MSA visualization application that focuses on simultaneous display of characteristics of the alignment and automatically downloaded metadata such as conserved domains, and allows the user to view the information at different levels of detail to enable easy recognition of interesting patterns for further analysis.

**CR Categories and Subject Descriptors:** H.5.2 [User Interfaces]: Graphical user interfaces (GUI), User-centered design; J.3 [Life and Medical Sciences]: Biology and genetics

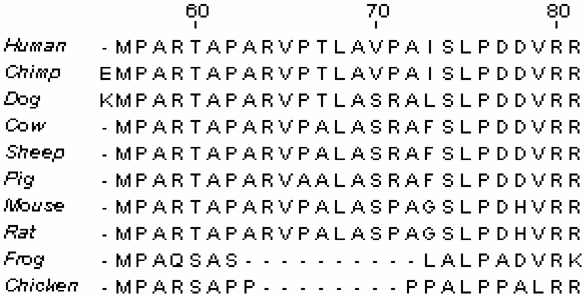
**Additional Keywords:** information visualization, bioinformatics, multiple sequence alignment, conserved domains

## 1 Introduction & Motivation

Proteins, DNA, and RNA are the basic molecules of life and each can be represented computationally as a sequence of characters. Of these, proteins are especially important because they are typically the molecules that carry out functional processes within cells. Proteins are made of chains of 20 different amino acids; therefore a protein sequence can be viewed as a string of characters from a 20 letter alphabet. For understanding proteins, multiple sequence alignment (MSA) has proven to be one of the most widely used bioinformatics methods because it allows biologists to analyze the similarities and differences of related proteins. Figure 1 shows a portion of a multiple sequence alignment of DNA methyltransferase proteins from 10 different organisms. The canonical representation of an MSA has each protein sequence on a separate line with matching characters aligned in columns and spaces inserted where necessary to improve the alignment. A series of one or more spaces is called a gap and is represented by dashes. The similarities and differences highlighted by multiple sequence alignments can lead to conclusions about the evolutionary history of the organisms, as well as information pinpointing functional parts of the sequences of each organism.

Biologists often want to investigate the "functional domains" of proteins. These are the sections of the protein's sequence that enable it to serve a particular biological role. Because these sections tend to be evolutionarily conserved (they remain the same in related organisms), they are also called conserved domains. After a sequence has been identified as a functional conserved domain experimentally or using predictive methods, a computer model of the domain can be generated (often using a Hidden Markov model). That model can be used to identify the domain in sequences from other organisms. There are many online databases that take a protein sequence as a query and return matching domains. For this application, the Conserved Domain Database (CDD) from NCBI [1] was used.

Jalview [2] is the tool that is most closely related to ours. Jalview displays an alignment, as well as many kinds of associated information including conserved domains. However, with Jalview, it is not possible to effectively view more than a few domains on an alignment, especially if the conserved domain matches overlap on a part of a sequence. Furthermore, it is not easily possible to get an overall view—across the entire alignment—of where each conserved domain lies.



```
             60        70        80
             |         |         |
Human   -MPARTAPARVPTLAVPAISLPDDVRR
Chimp   EMPARTAPARVPTLAVPAISLPDDVRR
Dog     KMPARTAPARVPTLASRALSLPDDVRR
Cow     -MPARTAPARVPALASRAFSLPDDVRR
Sheep   -MPARTAPARVPALASRAFSLPDDVRR
Pig     -MPARTAPARVAALASRAFSLPDDVRR
Mouse   -MPARTAPARVPALASPAGSLPDHVRR
Rat     -MPARTAPARVPALASPAGSLPDHVRR
Frog    -MPAQSAS----------LALPADVRK
Chicken -MPARSAPP--------PPALPPALRR
```

Figure 1. A view of a portion of an MSA in Jalview

Our application addresses this specialized problem by creating a new way to view an MSA concurrently with all conserved domains that match over the alignment.

## 2 Methodologies & User Interface

Our application was developed in Python, using wxPython for 2D graphics and BioPython for manipulating and managing the alignments and conserved domains.

Upon the input of a multiple sequence alignment file, the application queries the online NCBI CDD database for each sequence and parses the results. The first view presented is an overview of the alignment and the domains present as shown in Figure 2.

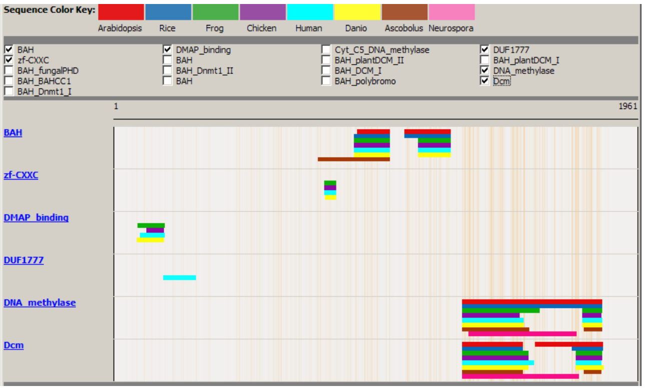
### 2.1 Alignment Overview

The initial overview alignment presents a condensed view of the entire alignment and associated conserved domains. In the example shown in Figure 2, the alignment spans amino acid indices 1 to 1709.

Unlike other systems for viewing alignments, this application gives an overview that initially focuses on each conserved

* e-mail: ADL91@msstate.edu
† e-mail: bridges@cse.msstate.edu
‡ e-mail: tjk@cse.msstate.edu

domain, rather than each sequence. The alignment is displayed separately for each domain and each domain is represented by a horizontal block. The background is a cream color whose saturation represents the strength of the alignment based on sum-of-pairs scoring at each position. A more saturated color represents a stronger match among the sequences at each column; a completely unsaturated color typically represents a column where many gaps were inserted.



☑ BAH
☑ zf-CXXC
□ BAH_fungalPHD
□ BAH_BAHCC1
□ BAH_Dnmt1_I
☑ DMAP_binding
□ BAH
□ BAH_Dnmt1_II
□ BAH
□ Cyt_C5_DNA_methylase
□ BAH_plantDCM_II
□ BAH_DCM_I
□ BAH_polybromo
☑ DUF1777
□ BAH_plantDCM_I
☑ DNA_methylase
☑ Dcm

Figure 2. The initial overview of an alignment with 6 conserved domains displayed.

Each sequence in the alignment is assigned a color, as shown in the key at the top of the screen. In the overview, a bar of color is drawn above the background whenever the respective conserved domain is present in that sequence across a specified portion of the alignment. In this example, it is easy to tell that the DNA_methylase domain (second from bottom) is present in every sequence near the end of the alignment. Similarly, the DUF1777 domain (fourth from top) near the beginning of the sequence is clearly only present in human. Furthermore, even though the last two domains overlap, it is still easy to see where they lie on the alignment since they are displayed on separate tracks.

This overview can be useful for drawing general conclusions about some parts of the alignment. Often, however, this view will motivate the user to look more closely at a specific part of the alignment. The application allows the user to click on the blue, underlined hyperlinks to view a single conserved domain's relationship on the alignment in greater detail.

## 2.2 More Detailed Views

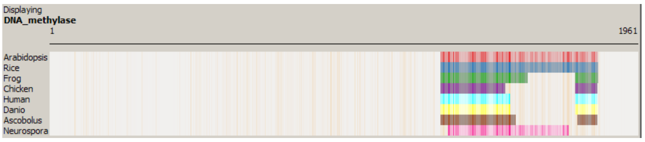
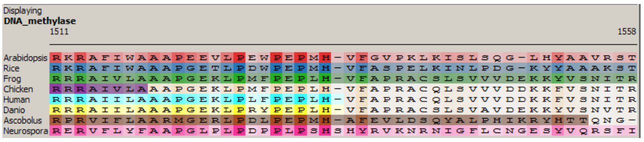


Figure 3. Overview of a single conserved domain.

This detail view appears below the overview, so that both are shown at the same time (a cropped view is shown in the figure). Initially, this view is very similar to the overview, in that it spans the entire alignment and has the same cream color background representing areas where the specific conserved domain is not present. However, in this view, the saturation of the colors that represent each sequence is adjusted in the same way the background is. This means the alignment strength can be visualized in conjunction the conserved domain. It may be important to compare the strength of the alignment in a domain.

To zoom in on a certain area of the alignment, a user can click and drag a rectangle around the area of interest. Upon releasing the mouse, the alignment window will redraw with only the selected area. When the user selects an area small enough to make drawing the individual amino acids characters viable, these are also displayed.



```
Arabidopsis RKRAFIWAAAPEEVLPEWPEPMH-VFGVPKLKISLSQG-LHYAAVRST
Rice        RKRAFIWAAAPGETLPDWPEPMH-VFASPELKINLPDG-KYYAAAKST
Frog        RRRAIVLAAAPGEKLPMFPEPLH-VFAPRACSLSVVVDEKKYVSNITR
Chicken     RRRAIVLAAAPGEKLPMFPEPLH-VFAPRACQLSVVVDDKKFVSNITR
Human       RRRAIILAAAPGEKLPLFPEPLH-VFAPRACQLSVVVDDKKFVSNITR
Danio       RRRAIILAAAPGEKLPRYPEPLH-VFAPRACSLSVAVDEKKYVSNVTR
Ascobolus   RPRVIFLAARMGERLPDLPEPMH-AFEVLDSQYALPHIKRYHTTQNG-
Neurospora  RERVFLYFAAPGLPLPDPPLPSHSHYRVKNRNIGFLCNGESYVQRSFI
```

Figure 4. Close-up view of the alignment showing individual amino acids.

Figure 4 shows a portion of the alignment where most of the sequences have the DNA_methylase domain present. In this view, a user can see each individual amino acid in every sequence, in addition to the strength of the matches represented by the saturation of the colors. In this example, it is easy to see that at position 1511 (the first position displayed), all of the amino acids match in every sequence, so the color is highly saturated. However, in the middle of the view as the alignment strength lessens (gaps are inserted and columns are mismatched), the colors become less saturated.

Users can move up and down the alignment using the scroll button on their mice, as well as the arrow keys on the keyboard. Additionally, users can click and drag the sequence labels on the left to move that sequence higher or lower in the stack, for closer analysis of a subset of the sequences, while still keeping information about all the sequences available.

If the user zooms in on a specific portion of the alignment, a selection box is drawn on the overview to represent which section of the alignment the user is viewing in relation to the whole alignment, as well as the other domains. If a user selects another domain, a new detail view will appear for that domain; additionally, the zoom factor will stay the same. Users can easily zoom back out to the full view by clicking the right mouse button.

## 3 Discussion & Conclusions

Our prototype application demonstrates a new method for displaying multiple sequence alignments and conserved domains at different levels of detail. We plan to investigate methods for displaying additional layers of information on the display, such as predicted DNA binding sites. We also plan to conduct user studies with biologists to evaluate the effectiveness of the display.

## Acknowledgments

This research was supported by an NSF EPSCoR award, grant number EPS-0556308.

## References

[1] Marchler-Bauer, A., Anderson, J. B., DeWeese-Scott, C., Fedorova, N. D., Geer, L. Y., He, S., Hurwitz, D. I., Jackson, J. D., Jacobs, A. R., Lanczycki, C. J., Liebert, C. A., Liu, C., Madej, T., Marchler, C. H., Mazumder, R., Nikolskaya, A. N., Panchenko, A. R., Rao, B. S., Shoemaker, B. A., Simonyan, V., Song, J. S., Thiessen, P. A., Vasudevan, S., Wang, Y., Yamashita, R. A., Yin, J. J. and Bryant, S. H. CDD: a curated Entrez database of conserved domain alignments. In *Nucleic Acids Research*, volume 31, 383-387, January 2003.

[2] Clamp, M., Cuff, J., Searle, S. M., Barton, G. J. The Jalview Java alignment editor. In *Bioinformatics*, volume 20, number 3, 426-427, February 2004.