MAVIS: VISUALIZATION TOOL FOR MULTIPLE SEQUENCE ALIGNMENTS

A Thesis

Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

By

Lei Zhao

May 2011

MAVIS: VISUALIZATION TOOL FOR MULTIPLE SEQUENCE ALIGNMENTS

Lei Zhao

APPROVED:

Dr. Yuriy Fofanov, Chairman
Department of Computer Science

Dr. Dan Graur
Department of Biology and Biochemistry

Dr. Shishir Shah
Department of Computer Science

Dean, College of Natural Sciences and Mathematics

Acknowledgements

(Under construction)

MAVIS: VISUALIZATION TOOL FOR MULTIPLE SEQUENCE ALIGNMENTS

An Abstract of a Thesis
Presented to
the Faculty of the Department of Computer Science
University of Houston

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

By
Lei Zhao
May 2011

Abstract

Biological sequences, such as DNA, RNA and protein, are frequently analyzed in the field of molecular biology and bioinformatics. Multiple sequence alignments (MSAs) help identify similarity between the sequences, and reveal their functional, structural or evolutionary relationships. MSA analysis often requires various visualization techniques, which assists researchers to better understand, evaluate and learn from the MSA results. One of the common techniques is the MSA coloring. There are several coloring tools available [1], but their color schemes focus more on chemical properties of the sequence residues, rather than the actual structural or evolutionary information of the whole alignment.

In this thesis, we introduce Mavis, a new approach to coloring MSAs and highlighting the quality and structure. Instead of using a pre-defined color scheme based on residue types, we designed a new algorithm to dynamically generate a color for each residue, based on a user-determined similarity scoring between them. This new tool colors an MSA in such a way that a well-aligned region is represented by a solid-color block, while a poor-aligned one by a mosaic with various colors. Thus, the alignment quality and internal structure are clearly displayed.

We implemented the core Mavis algorithm in R, the application programming interface (API) in Perl CGI, and several graphical user interfaces (GUIs) over web browsers and mobile devices like iPhone and iPad. We also followed common software design principles, and the system is extensible for even more add-on features and new user interfaces. The software is ready-to-use for any biologists without requiring deep computer science skills, and will deliver the visualization result of a typical MSA in minutes.

Contents

1 Introduction **1**

1.1 Background 1

1.2 Related Work 2

1.3 Thesis Outline 3

2 Approach **5**

2.1 Data Input 5

2.2 Color Generating 6

2.3 Color Optimization 7

2.4 MSA Visualization 7

3 Algorithms **8**

3.1 Multidimensional Scaling 8

3.2 Color Space Mapping 10

3.3 Hue Rotation and Optimization 11

3.4 Hue Flipping 13

3.5 Sequence Sorting 15

3.6 Time Complexity 16

4 Implementation **17**

4.1 R Script 18

4.2 API 20

4.3 User Interfaces 21

4.4 Design Principles 21

5 Results **22**

5.1 Mavis for Web 22

5.2 Mavis for iPhone 22

5.3 Mavis for iPad 22

5.4 Validity 22

5.5 Benchmarks 22

6 Discussion **24**

6.1 Color Spaces 24

6.2 Optimization Algorithms 26

References **28**

List of Figures

1.1 Example of Multiple Sequence Alignment 4

4.1 Example of Similarity Score File 18

4.2 Example of Coloring Result File 19

4.3 Example of API Response in JSON Format 20

List of Tables

6.1 Comparison of *L-BFGS-B* and *nlminb()* 27

Chapter 1

Introduction

1.1 Background

In the field of molecular biology and bioinformatics, sequence data, such as nucleic acid sequences (DNA and RNA) and protein sequences are widely used as research materials. To understand the sequence data in functional, structural or evolutionary aspects, multiple sequence alignment (MSA) analysis is required for almost all comparative sequence studies.

A multiple sequence alignment is usually a matrix, in which each row represents a sequence, and each column corresponds to the equivalent positions across all sequences aligned. Each element of the matrix contains a symbol, which can be either '-' (a gap) or a sequence symbol (an amino acid for DNA/RNA sequences, or a nucleotide base for protein sequences). [2]

Due to the size and complexity of the multiple sequence alignment datasets, computer visualization techniques are essential to assist in understanding, analyzing and evaluating the alignment results. Numerous MSA visualization applications, either stand-alone or web-based, have been developed in recent decades, ranging from simple MSA viewers, to complex editing and analysis platforms. Many of them have graphical user interfaces showing a grid of symbols, usually along with various coloring, sizing and annotation. [1]

1.2 Related Work

Coloring an alignment has been widely implemented to highlight specific regions, properties or patterns. [1] The simplest way of coloring is to use a fixed color scheme, in which each sequence symbol has its own color, and will not be changed across the alignment. The assignment of color schemes are usually based on some empirical properties or chemical classifications. There are numerous color schemes available for different purposes, but Taylor [3] and Clustal [4] are widely used and often considered de facto standards.

However, using pre-determined color schemes to color sequence symbols, either amino acids or nucleotide bases, are not always helpful to phylogenetic analysis. Those fixed schemes focus on chemical differences between specific types of symbols, rather than between sequences, regions, blocks or structures of the alignment. For example, if symbols are examined within a certain column, the colors reflect the similarities fairly well. However, if the alignment is analyzed horizontally, comparative information within a row or a rectangular region are unable to be displayed, like how one sequence is similar with others, which section aligns better than another, which subset of sequences in which range of columns are more conserved, or how robust the alignment accuracy is.

Here we suggest a new MSA visualization tool called Mavis (Multiple Alignment VISualization), which focuses on highlighting evolutionary relationships and internal structures of an alignment. Instead of relying on chemical properties or any other information from symbol types, our approach takes only symbol-symbol correlation into consideration. The most valuable result we would like to provide is a whole picture of the MSA.

1.3 Thesis Outline

This thesis is organized in the following way: Chapter 2 briefly summarizes how Mavis processes data input, creates colors and displays the colored MSA graph. Chapter 3 explains the algorithm step by step, from scaling to mapping and optimization. Chapter 4 describes our implementation of the algorithm, including the back-end, API and front-end. Chapter 5 presents test results to show the validity and performance of our implementation. The whole work is discussed and concluded in Chapter 6.

```
NC_001802.HXB2          M----QPIPI----------VAIVALVVAII-IAIVVWSI----VIIEYRKILRQR----KIDRLIDRLIER-AEDSG--NESEG---------------EISA-LVEMGVEMGHHAPWDV-DD-----------L
EF637049.B              M----QSLQI----------VAIVALVVTAI-IAIVVWSI----VLIEYRKLLRQR----KIDRLIDRIRER-AEDSG--NESEG------------DQEELAG-L----VERGHLAPWDV-DD-----------L
EF514700.B              M----QPLEI----------LAIVALVVAII-LAIVVWTI----VFIEYKKILRQR----KIDRLIDRIAER-AEDSG--NESEG------------DQEELSA-L----VDMGHDAPWVVVDQ------------
DQ056417.C              M------------LESIDYRLGVAALLLALI-IAIIVWII----AYLEYRKLLRQR----RIDKLIKRIRER-AEDSG--NESEG------------DIEELST-M----VDVEHLRLLDV-NN-----------L
AY463217.C              M---------VDLLAGVDYRVGVGALIIALI-IAIIVWIW----VYIEYRKLLRQR----KIDWLIKRLRER-EEDSG--NESEG------------DTEELAT-M----VDMGHLRLLDD-NN-----------V
DQ011165.C              M---------LNFLAGVDYRIGVGALIVGLI-IAIVVWII----VYLEYRKLVKQR----KIDWLIERIRER-AEDSG--NESEG------------DTEELAT-M----VDMGHLRLLDA-YD-----------L
AB254142.C              M---------INFAARVDYRVGVAAFTIALI-IAIVVWII----VYLE---LVRQR----KIDQLIIRIRER-EEDSG--NESEG------------DIEELST-M----VDMGQLRLLDG-NG-----------L
AY901969.C              MV---NLLEKVNLFEKVDYRLGVGALLIALV-IAIIVWTI----AYIEYRKLVRQR----KIDWLVKRIRER-AEDSG--NESDG------------DTEELST-M----VDLGHLRLLDV-AE-----------L
EU110088.A1             M----NQLQI----------LAIXGLVVALI-LAIVVWTI----VGIEYRKLLRQR----RIDRLIKRISER-AEDSG--NESDG------------DTEELSQ-L----VEMGNYNLGFD-DN-----------L
AB253428.A1             M----QLLEI----------CAVVGLVVALI-IAIVVWTI----VGIEYKKLLKQR----KIDRLVDRIRER-AEDSG--NESDG------------DREELSL-L----VDMGDYDLGDD-NN-----------L
AF457052.A1             ML---SALEI----------CAIAGLVIALI-IAIVVWTI----VGIEYRRLLKQR----KIDRLIERIRER-AEDSG--NESDG------------DTEELAA-L----IEMGNYDLGDA-ND-----------L
AF077336.F1             M----SYLLA----------IGIAALIVALI-IAIVVWTI----VYIEYKKLVRQR----KINKLYKRIRER-AEDSG--NESEG------------DAEELAA-L----GEMGPFIPGDI-NN-----------L
DQ168575.G              M----KSLEI----------SAIVGLIVAFI-AAIVVWTI----VLIEYRKIRKQK----RIDKILDRIRER-AEDSG--NESEG------------DTEELAT-L----VDMVDFEPW-VGDN-----------L
AY795907.D              M----QTLEI----------LSIVALVIAAI-IAIIVWTI----VYIEYRKIRRQR----KIDQLIDRIRER-AEDSG--NESEG------------DEEELST-L----MEMGHAAPWNVADD-----------L
EF394358.PTS            MVKI----------------VVGSVLTNVIGAFCILLILIGGGLL--IIAFVRRELERERQHQRVIERLVRRLSIDSGIDEDEEL------------NWNNFDPHN----------------FNP---RDW----I
AF103818.PTT            ML---NWFE-----------IGLIALGIEGI-LVVIIWGL----VARLWRQIKIKENTQQEIQNLLERIRIR-EEDSG--NESDGE-----------EEETLAK-L----LSSLEL------DNPRI--------V
DQ373066.PTT            MQ---LEI------------VLIILFIALML-VAIFAWIA----AYKEYKKLQQVR----RIERLQDRIRSR-AEDSG--NESDG------------DEILLVEEL----MQVHQHQVNPDWMDRILF-------W
DQ373063.PTT            MD---IVQQV----------GLLVVLIIELV-IVIVIWVK----VYKLCKEDRRQK----KIDRLIARIRER-AEDSG--NESDG------------DTEELQD-LITEGDNLMHIGIRDN-RN-----------N
DQ373065.PTT            ML---LLIK-----------LGFIGLAIETL-IVIVVWAI----VYRIYREVKVEE----KISQLRQRIRDR-AEDSG--NESDG------------DAEELAN-L----LPPDRI------DQDNW--------V
EF535994.PTT            ME---IFII-----------LGLIGIVIELV-IAIVVWLK----AYECYKALKRQE----RRDQLIDRIRER-AEDSG--NESDG------------DTEELEA-I----LIPEDRHVLVAIRG-----------Y
DQ373064.PTT            MD---L-IE-----------LGLIGLVIELI-IVIVVWLK----AYQLYKENIRQK----AINKLIERIRER-GEDSG--NESEG------------DMDELHA-I----LRSGDPE-LVLIDN------------
AY169968.PTT            MI---L--------------IALGCLAIALILLNIFIWRN----LWRLCKQNKLEV----EIENLALRLTER-AEDSG--NESDGG-----------EEEKLQR-LLHDHDNMHGF------ANPLFD-------I
EF535993.PTT            ML---TWEQ-----------IGLIALGIEGI-IATVVWGI----AFITWRRRKIEEDRAQKIIDLLERIRAR-QDDSG--NESDGE-----------DQRQLM---------AYGF------DNPMFD-------W
U42720.SIVCPXANT        MTNI----------------FEYAFL-----AFSIVLWIICIPILYKLYKIYKQQQIDNKRNQRIIEVLSRRLSIDSAIEEDEEA------------DTYYLGSGF----------------ANPVYREGD----E
AF382828.SIVCPXGAB2     ML---SMWVA----------IGLIGIGTLLV-INIVVWGIVGISVYKRWKRHKEEQ----RIIDLIIKITER-AEDSG--NESDGE-----------DKETLAT-L----LHNNGFDNPMFEDR-----------I
DQ017383.N              ML------E-----------LGFIAFGVAII-IAVIIWVL----LYKEYKEIKLQE----KIEKIRQRIRDR-AKESG--KESDGDAEWLGDKESVGDDEWLAI-L----LSPDKL------DN--W--------V
AY169807.O              ML---HRDL-----------LLLIIISALLL-TNIILWMF----VLRKYLEIKKQERREREILERLRXIREI-RDDSD--YESNEE-----------EEQEVRDHL----VHTFGF------ANPMFE-------I
AY169811.O              MH---HRDL-----------LTLIAVSALLF-INIILWIY----VLRKYLEQRKQDRREREILERLRRIXEI-GDDSD--YESNEE-----------EEQEVMD-L----VHNHGF------DNPMFE-------P
AY169804.O              MH---QRDL-----------LILIAVSILCL-ICILVWTF----NLRKYLEHRKQDKREREILERLRRVREI-RDDSD--YESBGE-----------EEQEVMD-L----IHSHGF------ANPLFE-------L
AJ302646.O              MH---HRDL-----------LALITTSALLL-TNVVLWTF----ILRQYLKQKKQDKREREILERLRRIRQI-EDDSD--YESDGT-----------EEQEVRD-L----VHSYGF------DNPMFE-------L
AY169803.O              MH---YRDL-----------STLIIVSALLL-INVXLWMF----ILRXYLEHKRQERREREILERLRRIREI-KDDSD--YESNGE-----------EEQEVMD-L----VHSHGF------DNPMFE-------L
AY169810.O              MN---YKEL-----------LSLIVVSVLLL-AAIVIWMF----ILKKYLEQKEQDRRERELLKRIERLXEX-RDDSD--YESNGD-----------EEQEVMH-L----VHTHGF------ANPMFE-------L
AY169815.O              MQ---HKDL-----------LILIITSALLL-INVILWLF----VLKQCLEQKKQTKREREIIRRLRRIREI-EDDSD--YESNGE-----------EEQTVRD-L----IHSHGF------DNPMFE-------L
AY169809.O              MQ---YKGL-------------LLIIIALLL-INVXVWMF----NLRKYLEQKKQERREREVINRLRRIREV-KDDSD--YESNGE-----------EEQEVME-L----VHSHGF------DNPMFE-------L
AY169806.O              MY---KDQI-----------ILIIIFCVVFL-IAACIWLF----ILKTYLEQKKQDRREKELLRRLQRIIEI-RDDSD--YESNGE-----------EEQEVMD-L----VHEHGF------VNPMFE-------L
FJ424864.GOR            MH---PRDI-----------IVIIIGITLLA-VTVIIWLK----IFALYLRDRRER----RFFDRLERLLSN-KEDEG--YES--------------NEEEAAE-L----MEMGNELGFDFNLH------------
AJ549283.CERCOPITHECUS  MN---TYWI-----------AAISVWCVIVLLAPCTLYLL-----YQSYKEHKRVSRFEQDFQRLVQXYQ---EXDSG--YEDEDE-----------DH---NS-F----------------DNPLFDDGDPDQWV
AY340701.CERCOPITHECUS  MN---YWWS-----------LVAITYSLILIALPVAAWAW-----WRYYKITKRFKRIDQEIQRLIQIHERR-RHDSGVDTESESE-----------QHEETHG-F----------------VNPVFND-DFGEWV
AF468658.CERCOPITHECUS  MSAAALWWW-----------GAAVITFIYFC-LAIFALYL----AWDKWIKGKPK-----IPVAVIRLVED--DEESGIFEDASS-------------EPNAY-GF----------------ANPGFE-------V
AJ580407.CERCOPITHECUS  MLKIK---------------LGPIEYVCLVF-AVVITWIAAIGVGYLAYRAYKSYREELRYI-----RLRLW-SIDSG--YESSQE-----------D-------------------------------------P
AY340700.CERCOPITHECUS  MN---YWYL-----------AAVIVTGIYFV-IAIFAFVL----AYQRWCKPKK------VEVSVIRLLEEG-DGDSGIFEDAEDD-----------MAESEHHAF----------------ANPAFE-------Q
AF468659.CERCOPITHECUS  MHPAAVWWW-----------GAAIITFIYLC-VALLALYL----AWDKWVKGKPKP----TQVAVIRLIED--EEDSGIYDDASS-------------ELTGFNGF----------------ANPGFE-------V
EF070331.SIVMUS01CM2500 MN---WWWF-----------AAAVVTAIYFV-IALVAFVL----AYQRWCQPQKGQ----VEVNVIRLLEEG-DTDSGIFEDAEDG------------SDDQRHGF----------------LNPAFE-------L
EF070329.SIVMUS01CM1246 MH---YWYL-----------GAAIVTAIYLV-IALVAFIL----AYQRWCQPKP-P----IEVNVLRLLEEG-DSDSGIFEDACDG-----------EDEDSHRAF----------------ANPSFE-------P
```

Figure 1.1: An example of multiple sequence alignment.

Chapter 2

Approach

In this thesis, we address a novel method to visualize multiple sequence alignments. The main objective is to correlate alignment similarity (or confidence) with color similarity. In the alignment matrix, a well-aligned area forms a solid-color shape, while a poor-aligned one becomes a mosaic of various colors. Alignment-level information, such as evolutionary relationships, phylogenetic characteristics and alignment accuracy should be revealed by examining the graph generated by our tool.

2.1 Data Input

Unlike most traditional coloring approaches, we start with an set of pairwise scores, rather than the alignment itself. Each of these scores represents the similarity or alignment confidence between two symbols aligned together in the same position, or, in the same column. Every pair of symbols in the same column has a score.

This set of data can be generated by GUIDANCE [5, 6], a measure quantifying alignment uncertainty. It produces a confidence score between 0 and 1, named GUIDANCE score, for each symbol, symbol pair, column and sequence in the alignment. The symbol-pair GUIDANCE confidence scores can be used in our study to visualize MSAs.

2.2 Color Generating

The colors of the symbols are generated in such way that the color differences represent symbol differences. We consider the confidence scores reversely, as distances, and convert all the scores of a column into a $N * N$ distance matrix, where N is the number of symbols. These N symbols can be represented by N points in at most $N - 1$ dimensions, such that the Euclidean distance between any pair of points is exactly equal to the distance in the matrix. Once we can map these points to a color space, those corresponding colors should be able to fulfill our requirements.

However, most color spaces are three-dimensional or under. It is necessary to reduce the dimension from $N - 1$ to at most 3 in order to process the mapping, but still preserve the distances as much as possible. This dimension reduction is done by a statistical technique called *classical multidimensional scaling* [7].

2.3 Color Optimization

Up to now, our approach only operates on the column level. Further adjustment is still needed to achieve the alignment-level objectives. We notice that the color space can be arbitrarily rotated and/or flipped, without changing the distance between any pair of symbols. For instance, all reds and all greens are virtually identical in the sense of distances and quality of alignment. Converting them into the same color pattern will better support this idea and remove unnecessary color noises.

We defined a penalty function to quantify the overall noise level of an alignment. The ideal combination of rotating and flipping each column should result in the lowest penalty value. For large MSAs, finding the global optimum is not always possible, and the approximate solution is also an acceptable option.

2.4 MSA Visualization

We choose to implement a web-based user interface, because of its flexibility, portability and light weight. Many features are provided to enhance the visualization effects, such as sorting sequences by their average colors, toggling to show/hide symbol names, and multiple-MSA support. An API between the front-end and the back-end is designed to make the system more extensible.

Chapter 3

Algorithms

The algorithm of color computation consists of five stages. First, multidimensional scaling reduces the number of dimensions and convert the confidence/similarity scores into coordinates. Second, coordinates are mapped onto part of the *CIE Lab color space* [8], and colors are created. Third, colors are flipped and rotated to smooth the graph and remove color noises. Finally, calculate the average color/hue of each sequence, sort them, and output for further display.

(Flowchart)

3.1 Multidimensional Scaling

This stage begins with the confidence scores described above. Each column in the MSA has a complete set C of pairwise scores. Each pair of non-gap symbols i, j in this column has a score $c_{i,j} = c_{j,i} \in [0,1]$. Here 1 stands for absolute confidence or similarity, while 0 for the opposite.

This dataset can be easily converted to a set D of dissimilarities or distances between two symbols, by

$$d_{i,j} = d_{j,i} = 1 - c_{i,j}$$

These distance values are re-organized as an $n \times n$ symmetric distance matrix M

$$\begin{array}{cccccc}
 & 1 & 2 & 3 & ... & n \\
1 & 0 & d_{1,2} & d_{1,3} & \cdots & d_{1,n} \\
2 & d_{2,1} & 0 & d_{2,3} & \cdots & d_{2,n} \\
3 & d_{3,1} & d_{3,2} & 0 & \cdots & d_{3,n} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
n & d_{n,1} & d_{n,2} & d_{n,3} & \cdots & 0
\end{array}$$

where n is the number of non-gap symbols in the current column.

These n symbols are placed into a two-dimensional space, in which the distances between each other are still preserved as much as possible. This procedure involves a statistical technique called *classical multidimensional scaling (classical MDS)*, also known as *(*Torgerson-Gower scaling) or *principle coordinates analysis (PCoA)* [9]. Multidimensional scaling takes a distance matrix and assigns a coordinate to each item in a lower dimensional space, such that the Euclidean distance between any two coordinates is approximately equal to the original distance value between the corresponding items.

An statistical function called *cmdscale()* [10] is utilized in this multidimensional scaling procedure. This function is written in R, which is a programming language for statistical computing [11], and is provided in Rs stat package. It reads a distance matrix and returns a set of points in a k-dimensional space, where k is a user-defined parameter to the function, and must be less than the number of points n. [12, 13]

In our approach, k is set to 2, meaning the distance matrix is scaled down to a two-dimensional space. However, a column of an MSA may have as few as one or two non-gap symbols, so that the distance matrix of this column is only 1×1 or 2×2. In such cases, $k = 2$ will not be a valid parameter to *cmdscale()* because of the $k < n$ requirement. So we let

$$k = \begin{cases} 0 & \text{if } n = 1 \\ 1 & \text{if } n = 2 \\ 2 & \text{otherwise} \end{cases}$$

Then we perform the multidimensional scaling on the distance matrix M. When $n < 2$, the generated coordinates are lower than two-dimensional, and the missing coordinate values are filled by zeros.

(Flowchart)

3.2 Color Space Mapping

The above procedure returns a set P of n coordinates (x, y) in a two-dimensional space, representing the n symbols in a column of a multiple sequence alignment. This two-dimensional space is further mapped to a three-dimensional color space.

The CIE Lab color space [8] is a color-opponent space. The name *Lab* stands for the three dimensions: L for lightness, and a and b for the color opponents based on nonlinearly compressed *CIE XYZ color space* coordinates [14, 15]. Compared to the *RGB* and *CMYK* color models, the design of Lab color emphasizes more on approximating human vision rather than physical devices. This feature makes it suitable for our visualization purpose. [16]

Coordinates in P have only two dimensions x and y. To map P to the three-dimensional CIE Lab space, L for lightness is set to a fixed value 75, and two other dimensions a and b are assigned as x and y scaled from $[0, 1]$ to $[0, 100]$.

$$a = x * 100$$

$$b = y * 100$$

At the end of this step, n CIE Lab colors (L, a, b) are created.

(Flowchart)

3.3 Hue Rotation and Optimization

One of the limitations in many previous MSA visualization techniques is, they introduced unnecessary confusion by using a fixed color scheme. Even for exactly same aligned columns, colors may be totally different, which won't represent any biological dissimilarities. To minimize this effect, we convert the colors from CIE Lab space to *CIE LCH space*, and perform rotations to eliminate noisy colors to the greatest extent possible. In this way, we smooth the color pattern and emphasize the real similarities and dissimilarities among sequences and blocks.

The *CIE LCH color space*, also known as polar-Lab color space, is a cylindrical transformation of the CIE Lab space so that the a and b axles are converted to a polar coordinate system. The radial coordinate C measures *chroma* and the angular coordinate H measures *hue*. LCH space makes it easier to rotate colors, by changing their *hue* value. The CIE LCH colors can be converted from CIE Lab colors using an R library called *colorspace* [17, 18].

Hue values are circular, meaning 0 and 2π are identical. Given two angles p and q, the difference between them is given by

$$\operatorname{diff}(p, q) = (p - q + \pi) \bmod 2\pi - \pi$$

This calculation returns the difference in the interval $[-\pi, \pi)$, or in degrees $[-180°, 180°)$. The absolute value of the difference

$$|\operatorname{diff}(p, q)| = |(p - q + \pi) \bmod 2\pi - \pi|$$

returns in the interval $[0, \pi]$, or in degrees $[0°, 180°]$.

Given a set A of angles a_i, the average angle $\overline{A}$ is

$$\overline{A} = \arctan \frac{\sum\limits_{i=1}^{n} \sin a_i}{\sum\limits_{i=1}^{n} \cos a_i}$$

and the variance $\operatorname{Var}(A)$ is

$$\operatorname{Var}(A) = \frac{\sum\limits_{i=1}^{n} \operatorname{diff}(\overline{A} - a_i)^2}{n}$$

Given a colored MSA Z with m rows and n columns, to estimate its overall noise level, a penalty function $\text{pen}(Z)$ is defined by

$$\text{pen}(Z) = \sum_{j=1}^{m} \text{Var}(h_{1j}, h_{2j}, \cdots, h_{nj})$$

where h_{ij} is the hue value of the symbol at the ith column and jth row.

We look for an optimized set R of n rotation angles r_i, such that after rotating each hue in every ith column of Z clockwise by r_i, the rotated MSA Z' achieves the minimized return value of the penalty function $\text{pen}(Z')$.

The optimization procedure is performed by R's general-purpose optimization function *optim()* [10]. Among five available optimization algorithms provided by this function, we choose the *bounded Broyden-Fletcher-Goldfarb-Shanno* (L-BFGS-B) method [19]. This algorithm allows box constraints, meaning each variable can be given a lower and/or upper bound. Plus, it uses a limited amount of memory. The initial values are set to 0 and boundaries to $[0, 2\pi]$. More detailed comparison between different optimization functions and algorithms will be discussed later in Chapter 6.

(Flowchart)

3.4 Hue Flipping

One characteristic of the coordinates generated from multidimensional scaling is, they can not only be rotated, but also flipped, without changing the distances between each other. Adding flipping ability to the above rotation optimization procedure may result in lower minimal penalty values and better coloring results. To keep the penalty function simple and fast, we perform a heuristic flipping *before* the rotation step.

For each pair of adjacent columns C_i and C_{i+1}, $1 \leq i \leq n-1$, compare their hues row by row. Rows with gaps in these two columns are not take into consideration. For those rows with non-gap symbols on both sides, record their hue values

$$h_{i,1}, h_{i+1,1}, h_{i,2}, h_{i+1,2}, \cdots, h_{i,k}, h_{i+1,k}$$

into two lists H_i and H_{i+1}:

$$H_i = h_{i,1}, h_{i,2}, \cdots, h_{i,k}$$

$$H_{i+1} = h_{i+1,1}, h_{i+1,2}, \cdots, h_{i+1,k}$$

In both lists, compare two neighboring hue values $h_{i,j}$ and $h_{i,j+1}$, $1 \leq j \leq k-1$, using diff() function described in section 3.3. If $\text{diff}(h_{i,j}, h_{i,j+1}) < -\frac{\pi}{8}$, the change from $h_{i,j}$ to $h_{i,j+1}$ is called a *clockwise change*; if $\text{diff}(h_{i,j}, h_{i,j+1}) > \frac{\pi}{8}$, the change is called a *counterclockwise change*; otherwise, it's considered to be unchanged.

If a column contains more clockwise changes than counterclockwise changes, it's called *clockwise column*; if it contains more counterclockwise changes, it's called *counterclockwise column*; otherwise, it's a neutral column. Note that since gap rows are filtered out of the comparison, the type of a column may change when compared to different adjacent columns. For instance, a column may be clockwise compared to the previous column, and counterclockwise compared to the next column.

If two adjacent columns are in opposite types, flip the latter one to make them in same type. Then the flipped alignment are sent back to the procedure discussed in section 3.3.

(Flowchart)

3.5 Sequence Sorting

The alignment graph is supposed to provide information on how sequences will group with each other. We sort sequences after coloring optimization, based on the average hue of each one, which is calculated by the average angle function proposed previously in section 3.3.

The fact that hue values are circular, results in an additional question on where to cut the circle to get a sorted list of sequences. In a sorted circle of n average hue values, $h_1, h_2, \cdots, h_n$, we calculate the absolute difference between each pair of adjacent ones (h_1 is adjacent to h_n).

$$\Delta h_{1,2} = |\mathrm{diff}(h_1, h_2)|$$

$$\Delta h_{2,3} = |\mathrm{diff}(h_2, h_3)|$$

$$\cdots$$

$$\Delta h_{n,1} = |\mathrm{diff}(h_n, h_1)|$$

Then, the circle is split at the biggest gap

$$\Delta h_{max} = \Delta h_{i,i+1} = |\mathrm{diff}(h_i, h_{i+1})|$$

to best avoid separating similar sequences.

(Flowchart)

3.6 Time Complexity

Chapter 4

Implementation

We implemented the entire visualization pipeline on a Mac OS X server. The implementation consists of an R script as back-end, a Perl CGI script as API, and an HTML/JavaScript web page as user interface.

The R script *coloring.r* takes the pair-wise similarity score file as input, performs multidimensional scaling and color optimization, sorts sequences, and output colors and sequence information into two tab-separated files *colors.tsv* and *seqinfo.tsv*. The front-end web page takes an alignment ID as parameter and sends it to the API. The Perl CGI script parses the color file and sequence information file, along with the FASTA-format alignment file, and sends them back to front-end in JSON format [20]. The front-end uses JavaScript to generate the alignment matrix and render it with colors and symbol letters.

4.1 R Script

The script *coloring.r* begins with reading three arguments from command-line calls: score file name, colors file name and sequence information file name. The first file is for input and the latter two are for output. An example of calling this script on UNIX operating system is given below.

```
Rscript coloring.r data/score.tsv data/colors.tsv data/seqinfo.tsv
```

The file *score.tsv* contains four columns separated by tabs: column number, first row number, second row number, and similarity score between these two rows in this column. The row and column numbers start from 1, and the similarity score values range from 0 to 1. Figure 4.1 shows a few example lines in a *score.tsv* file.

```
#COL_NUMBER   #ROW_NUMBER_1   #ROW_NUMBER_2   #RES_PAIR_HIT/NALT
1             1               2               1.000000
1             1               6               0.933333
1             1               10              0.866667
1             2               6               0.933333
1             2               10              0.866667
1             6               10              0.133333
2             1               6               1.000000
2             1               10              1.000000
2             6               10              1.000000
```

Figure 4.1: An example of similarity score file *score.tsv*, showing the format of the file. The four columns are: column number, the first row number, the second row number, score value.

The script *coloring.r* consists of four parts. The first part reads the input file, performs multidimensional scaling and create color for each symbol in the CIE LCH space. The second part rotates and flips color hues and optimize the penalty function. The third part sorts the sequences by their average hues and calculates each sequences average RGB color. The last part outputs the symbol colors and sequence information into corresponding files.

The color file *colors.tsv* has five columns separated by tabs: column number, row number, and the symbol color in RGB triplet (red, green, blue) in the range from 0 to 1, which will be converted into hexadecimal format used in the user interface module. The sequence information file seqinfo.tsv also has five columns: row number, average hue from 0 to 360 degrees, and average color in RGB triplet. Figure 4.2 shows how a typical *color.tsv* file looks like.

```
1       1       0.1065524       0.5616907       0.3943046
1       2       0.106256        0.5616652       0.3936842
1       6       0.1065487       0.5616904       0.3942968
1       10      0.1065581       0.5616912       0.3943165
2       1       0.09057406      0.5677889       0.2867890
2       2       0.09057406      0.5677889       0.2867890
2       6       0.09057406      0.5677889       0.2867890
2       10      0.09057406      0.5677889       0.2867890
3       1       0.07634505      0.5664451       0.2581746
3       6       0.1113592       0.570106        0.3250781
3       10      0.07634505      0.5664451       0.2581746
```

Figure 4.2: An example of coloring result file *color.tsv*, showing the format of the file. The five columns are (from left to right): row number, average hue, red, green and blue.

4.2 API

A Perl CGI script *api.pl* and a Perl package *Mavis::API* act as the API between the R back-end and HTML/JavaScript front-end. The Perl CGI *api.pl* takes two parameters from HTTP requests, *action* and *id*. The default action is *alignment*, meaning the alignment information, including sequences, colors, average colors and order of sequences, should be returned. In this action, an additional *id* parameter is expected as the alignment ID. Another available action is *list*, which requests a list of all existing alignment IDs.

The API returns the query result to the web page in JSON (JavaScript Object Notation) format, which is a lightweight text-based data-interchange standard. JSON is derived from a subset of JavaScript syntax and can be easily parsed in JavaScript as well as many other languages, like Perl. We included the Perl JSON 2.51 module to provide JSON encoding/decoding ability to our API. [20]

Figure 4.3 is an example of the API response in JSON format:

```
{ "status"    : 1,
  "id"        : "demo",
  "colors"    : [ ["#666","#666","#666","#666"],
                  ["#665","#667","#667","#665"]],
  "sequences" : [ {"name":"Seq02", "color":"#26e", "seq":"DEMO"},
                  {"name":"Seq01", "color":"#26e", "seq":"DEMO"}]
}
```

Figure 4.3: An example of Mavis API response in JSON format.

4.3 User Interfaces

We implement an web interface with HTML and JavaScript to send API requests and draw visualization graphs. The JavaScript code dynamically communicates with API, creates an HTML table, and fills in colors and symbols. This part is simplified by jQuery, which is a popular cross-platform JavaScript library that abstracts complicated DOM selection, CSS manipulation and Ajax effects.

4.4 Design Principles

Chapter 5

Results

5.1 Mavis for Web

5.2 Mavis for iPhone

5.3 Mavis for iPad

5.4 Validity

5.5 Benchmarks

To evaluate the validity and performance of our approach, we tested it in two different ways. First, we run it on a few artificial datasets, to see if it creates the color patterns as we expected. Then, we run it on the BAliBASE dataset, observe how alignment size, in both number of rows and columns, will affect the excution time, and compare it with other tools.

Chapter 6

Discussion

In this report, we introduced a new way of coloring and visualizing multiple sequence alignments based on symbol-symbol confidence (or similarity) scores. Our main objective has been to create an intuitive way of showing the quality and internal structure of an alignment, using colors, which is the most natural and sensible way for human eyes. The basic idea is, the greater the dissimilarity between sequences, the more obvious difference in colors.

6.1 Color Spaces

Most previous techniques were based on fixed color schemes, meaning each sequence symbol, like an amino acid or a nucleotide, is assigned to a pre-selected color. This is of course a simple and straightforward solution, and is easy for observers to find a particular symbol or pattern. However, this fixed-scheme approach does not emphasize the relationship between adjacent columns and the internal regions and structures in the level of the whole alignment. This is the main reason why we don't use any predetermined color scheme, but calculate colors only based on the distance matrices, and rotate and flip colors to make them as smooth as possible. This strategy brings significant improvement to the alignment coloring, by showing the better-aligned regions and blocks in same or similar colors, while those irrelevant ones in quite different colors.

Since we are to convert the scaled coordinates to colors, choosing a suitable color space is an fundamental task. At first we scaled the distance matrix down to a three-dimensional space and mapped it directly to the RGB color space. However, we soon found a problem, that some colors, like very dark ones and very light ones, did not perfectly serve the purpose of representing distances, yet made the graph more noisy. We realized that we don't really need the whole color space and all possible colors. Instead, those colors with proper range of lightness and different hues will be enough to do the job. So we chose to scale down to a two-dimensional space and map it to CIE Lab space with a fixed lightness value (75). The reason why we didnt choose one-dimensional scaling is that, a linear space either could not map to enough number of colors (for example, use only one primary color, like red), or could not preserve the distance information (for example, use only hues). So two dimensions is a good balance.

6.2 Optimization Algorithms

In R, there are several general purpose optimization packages which offer facilities for solving our color rotation and flipping problems. Two popular functions are optim() and nlminb() from package stats. Function optim() provides implementations of five algorithms: *Broyden-Fletcher-Goldfarb-Shanno (BFGS)*, *bounded BFGS (L-BFGS-B)*, *conjugate gradient (CG)*, *Nelder and Mead (Nelder-Mead)*, and *simulated annealing (SANN)*. *Nelder-Mead* [21], which is the default one, returns robust results but is relatively slow. *CG* [22] in faster in larger optimization problems, but more fragile. *BFGS* is a balance, and *L-BFGS-B* further provides the ability of box constraints, that each variable can be given a lower/upper bound. *SANN* [23] is more powerful on rough surfaces but relatively slow. Another function *nlminb()* offers similar box constraint optimization and similar performance to *L-BFGS-B*, so these two algorithms are chosen to a further test.

Optimization algorithms always suffer from the local versus global minimum problem, and the final result are more or less unstable and depending on the initial values. To decide which one of *L-BFGS-B* and *nlminb()* is more stable in our approach, we run a test on both of them. The test dataset is the alignment of 44 Vpu protein sequences from GUIDANCE server. [6] We randomly created 100 sets of initial values, performed optimizations, and see how the return value of the penalty function (described in section 3.3) changed. Table 6.1 shows the results of the comparison. Obviously *L-BFGS-B* is more stable in our algorithm.

Table 6.1: Comparison of *L-BFGS-B* and *nlminb()*

Algorithm	Min. Penalty	Avg. Penalty	Max. Penalty	Std. Deviation
Initial	370,584	425,728	444,771	13,998
L-BFGS-B	136,909	178,850	247,218	17,871
L-BFGS-B (3 times)	136,909	178,850	247,218	17,871
mlninb()	146,759	181,688	426,535	34,579
mlninb() (3 times)	142,888	171,380	325,469	21,402

References

[1] J. Procter, J. Thompson, I. Letunic, C. Creevey, F. Jossinet, and G. Barton, "Visualization of multiple alignments, phylogenies and gene family evolution," *Nature Methods*, vol. 7, no. 3 Suppl, pp. S16–25, 2010.

[2] R. C. Edgar and S. Batzoglou, "Multiple sequence alignment," *Current Opinion in Structural Biology*, vol. 16, no. 3, pp. 368–373, 2006.

[3] K. Lin, A. C. W. May, and W. R. Taylor, "Amino acid encoding schemes from protein structure alignments: multi-dimensional vectors to describe residue types," *Journal of Theoretical Biology*, vol. 216, no. 3, pp. 361 – 365, 2002.

[4] J. D. Thompson, T. J. Gibson, and D. G. Higgins, "Multiple sequence alignment using ClustalW and ClustalX.," *Current Protocols in Bioinformatics*, p. 2.3, 2002.

[5] O. Penn, E. Privman, G. Landan, D. Graur, and T. Pupko, "An alignment confidence score capturing robustness to guide tree uncertainty," *Molecular Biology and Evolution*, vol. 27, no. 8, pp. 1759 –1767, 2010.

[6] O. Penn, E. Privman, H. Ashkenazy, G. Landan, D. Graur, and T. Pupko, "GUIDANCE: a web server for assessing alignment confidence scores," *Nucleic Acids Research*, vol. 38, no. Web Server, pp. W23–W28, 2010.

[7] I. Borg and P. J. F. Groenen, *Modern multidimensional scaling: theory and applications*. Springer, 1997.

[8] K. McLaren, "XIII—The development of the CIE 1976 (L* a* b*) uniform colour space and colourdifference formula," *Journal of the Society of Dyers and Colourists*, vol. 92, no. 9, pp. 338–341, 1976.

[9] J. C. Gower, "Some distance properties of latent root and vector methods used in multivariate analysis," *Biometrika*, vol. 53, no. 3-4, pp. 325–338, 1966.

[10] R Development Core Team, *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[11] R. Gentleman and R. Ihaka, "R: a language for data analysis and graphics," *Journal Of Computational And Graphical Statistics*, vol. 5, no. 3, pp. 299–314, 1996.

[12] F. Cailliez, "The analytical solution of the additive constant problem," *Psychometrika*, vol. 48, no. 2, pp. 305–308, 1983.

[13] M. A. A. Cox and T. F. Cox, "Multidimensional scaling," in *Handbook of Data Visualization*, pp. 315–347, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[14] C.I.E., *Commission internationale de l'Eclairage proceedings, 1931.* Cambridge University Press, Cambridge, 1932.

[15] T. Smith, "The C.I.E. colorimetric standards and their use," *Transactions of the Optical Society*, vol. 33, no. 3, pp. 73–134, 1931.

[16] D. Margulis, *Photoshop LAB color: The canyon conundrum and other adventures in the most powerful colorspace.* Peachpit Press Berkeley, CA, USA, 2005.

[17] R. Ihaka, P. Murrell, K. Hornik, and A. Zeileis, *colorspace: color space manipulation*, 2009. R package version 1.0-1.

[18] A. Zeileis, K. Hornik, and P. Murrell, "Escaping RGBland: selecting colors for statistical graphics," *Computational Statistics & Data Analysis*, vol. 53, pp. 3259–3270, 2009.

[19] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, vol. 16, p. 1190, 1995.

[20] D. Crockford, "The application/json media type for javascript object notation (json)," tech. rep., RFC 4627, July, 2006.

[21] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308 –313, 1965.

[22] R. Fletcher and C. M. Reeves, "Function minimization by conjugate gradients," *The Computer Journal*, vol. 7, no. 2, pp. 149 –154, 1964.

[23] C. J. P. Bélisle, “Convergence theorems for a class of simulated annealing algorithms on rd,” *Journal of Applied Probability*, vol. 29, no. 4, pp. 885–895, 1992.