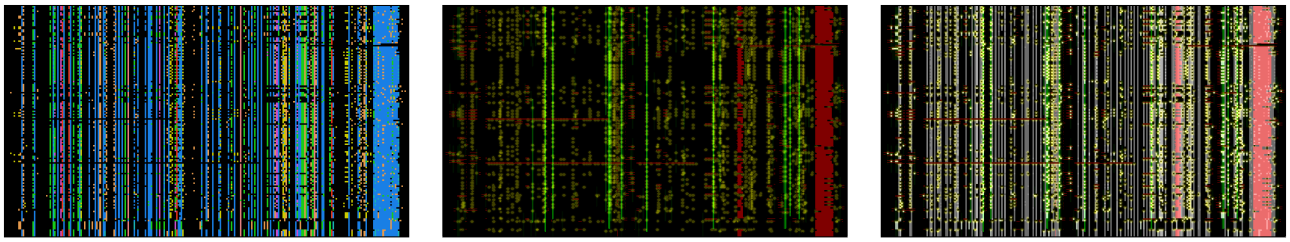
Large-Scale Multiple Sequence Alignment Visualization through Gradient Vector Flow Analysis

Khoa Tan Nguyen* Timo Ropinski†

Scientific Visualization Group, Linköping University, Sweden



(a) Clustal color scheme (b) GVF-based MSA features (c) Visual annotation

Figure 1: Results of the proposed visualization technique applied to the L1R_F9L family. While the visualization using the Clustal color scheme provides an overview, features such as absolutely or highly conserved areas are hard to detect (a). By using GVF analysis, we can extract conserved areas (green), gaps inserted by the alignment algorithms (red), and areas containing residues of low conservations (yellow) (b). The detected features can be imposed onto the original MSA (here shown in gray scale) to support visual annotation (c).

ABSTRACT

Multiple sequence alignment (MSA) is essential as an initial step in studying molecular phylogeny as well as during the identification of genomic rearrangements. Recent advances in sequencing techniques have led to a tremendous increase in the number of sequences to be analyzed. As a result, a greater demand is being placed on visualization techniques, as they have the potential to reveal the underlying information in large-scale MSAs. In this work, we present a novel visualization technique for conveying the patterns in large-scale MSAs. By applying gradient vector flow analysis to the MSA data, we can extract and visually emphasize conservations and other patterns that are relevant during the MSA exploration process. In contrast to the traditional visual representation of MSAs, which exploits color-coded tables, the proposed visual metaphor allows us to provide an overview of large MSAs as well as to highlight global patterns, outliers, and data distributions. We will motivate and describe the proposed algorithm, and further demonstrate its application to large-scale MSAs.

1 INTRODUCTION

Sequences of DNA, RNA, and proteins are the fundamental components of modern biological research. Multiple sequence alignment (MSA) is a way to organize these sequences, such that similar sequence features can be identified [21]. With respect to protein sequence analysis, a feature can be for instance the protein's structure, its function, or its homology with an ancestor. In the context of genomic analysis, insertion places for engineered gene sequences are commonly investigated features of interest.

Recent advances in the development of sequence alignment algorithms have led to an improved performance for large data sets, which enables domain experts to generate large-scale MSA data. While MSA accuracy has also been improved, no single alignment algorithm is perfect for all data sets [16], and MSAs from even the most accurate MSA algorithm can contain errors, often referred to as alignment artifacts [16]. Due to the increasing number of biological modeling methods depending on MSAs, it is essential that domain experts can bring in their knowledge to improve the data or to discover features of interest. For these purposes, a variety of tools and algorithms have been proposed for MSA visualization within the last two decades [22], of which many exploit the use of color-coding schemes designed for a better feature emphasis. Although color-coding schemes are known to provide an efficient visual metaphor for representing MSAs, they can also cause confusion in regions with complex patterns [17]. In addition, the sole use of color-coding schemes limits the effectiveness when dealing with large-scale MSAs. As modern MSAs can easily incorporate ten-thousands of sequences, it is clear that showing an overview as a color-coded table is not appropriate to enable the discovery of previously unknown features.

To support the analysis of large-scale MSA data sets, we propose a novel visualization approach that emphasizes features which are of potential interest. We transform the traditional table-based representation of the MSA data into pixel-based representations, which we visually annotate to enable exploration and discovery. Our approach is motivated by three reoccurring analysis goals which we have identified in cooperation with domain experts who frequently analyze MSA data. First, the exploration of regions with different levels of conservation, which helps to identify residues that are important or even critical, depending on the degree of their conservation. Second, areas containing changes caused by the alignment process need to be detectable. Third, the detections of regions where there is no observable conservation are of interest in the context of sequence engineering. For instance, when applying rational protein design, these regions can help to identify locations for site-directed mutagenesis. While the remainder of this paper focuses mainly on protein sequence analysis, the set of goals is often similar for genome sequence analysis, and our approach can therefore be extended to apply there.

In a traditional table-based representation of a MSA, there is

*e-mail:tan.khoa.nguyen@liu.se
†e-mail:timo.ropinski@liu.se

IEEE Symposium on Biological Data Visualization 2013
October 13 - 14, Atlanta, Georgia, USA
978-1-4799-1659-7/13/$31.00 ©2013 IEEE

a strong downward transition of the color patterns in highly conserved areas. On the other hand, the color patterns change disruptively or present sideway transitions in the areas containing gaps inserted by alignment algorithms. Consequently, to support the identification of absolutely or highly conserved areas as well as no observable conservation areas, which helps to target the mentioned goals, we employ gradient vector flow (GVF) analysis. GVF analysis enables us to detect relevant patterns, while at the same time supporting their shape aware highlighting for visualization purposes. The latter is of importance as feature extraction and annotation should not disturb the overall structure of the underlying MSA data. To be able to apply GVF analysis to MSA data, we transform a pixel-map of the MSA data into a vector field representation which we generate by taking into account the principal conservation axis, e. g., typically the y-axis as used in standard table-based MSA visualizations. Therefore, we can compute similarity and dissimilarity vectors to which we apply GVF analysis to extract relevant features on a larger scale and to further support their visual emphasis. The thus achieved visual representation has been designed by taking into account both, the capabilities of the human visual system and the goals identified together with domain experts. As a result, we are able to visually annotate the existing table-based MSA visualizations in order to visually guide domain experts during the exploration process. Our two main contributions are:

- A novel approach to identify patterns within large-scale MSA data, based on gradient vector flow analysis.
- An intuitive visualization metaphor, which we have designed to enable the visual annotation of large-scale MSA visualization, and thus support visual guidance.

The remainder of the paper is structured as follows. In the next section, we review work related to our approach. In Section 3 we describe our novel MSA feature extraction approach exploiting GVF analysis, while Section 4 introduces the proposed visualization approach developed to highlight the extracted features of interest within large-scale data. We present the results of the proposed visualization technique applied to different MSA datasets in Section 5, and conclude the paper in Section 6.

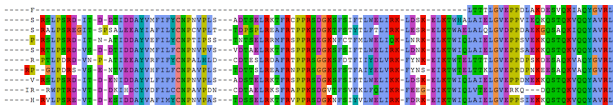
2 Related Work

MSA visualization. Many tools for MSA visualization have been developed over the past 20 years [22]. The output of an alignment algorithm can be interpreted as a matrix, in which each row is a sequence and each column defines equivalent positions across all sequences. Consequently, conventional MSA visualization techniques have adopted a table-based approach [2, 9, 10]. For instance, sequences are represented as rows and the corresponding residues and bases are depicted as letters arranged on a grid. Despite its simplicity, this type of visual representation provides a good control over the automatic generation of figures as well as the ability to work in both desktop and web-based environments. To further improve the visualization of MSAs, various color-coding schemes have been proposed [1, 17, 33]. The incorporated coloring effect helps to identify regions where specific properties predominate and at the same time highlight the variations. The existing schemes can be classified into two main categories: quantitative schemes which convey trends in specific empirical properties, and qualitative schemes which depict general physiochemical classes. Among various color-coding schemes provided in different visualization systems, the Clustal [17] and the Taylor scheme [18] are widely supported, and used [22]. Larkin et al. [17] also proposed an alternative way of using coloring schemes called *shading* to avoid confusion in regions with complex patterns of variations. Specifically, symbols are shaded on the basis of both their type and their predominance at each alignment position.

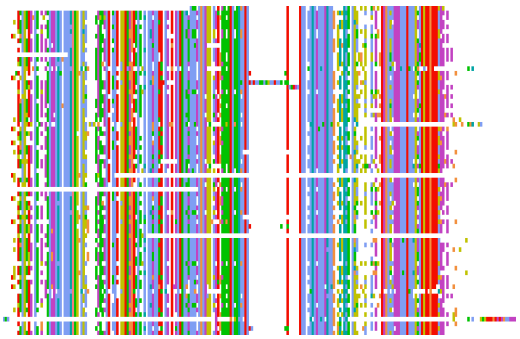
When dealing with the visualization of large-scale MSAs, there are two major approaches: clustering of sequences and profile-based representation. The common goal of these two approaches is to reduce the amount of data to be visualized. By applying frequency analysis, Schneider et al. [27] transform the traditional table-based representation of the MSA into a frequency-based profile representation. The authors proposed a compact visual representation of the input MSA that reveals not only the consensus sequence but also the relative frequency of bases and the information content at every position in a site or sequence. Schuster-Böckler et al. [28] proposed an extension to the Sequence Logos visual representation by incorporating both emission and transition of probabilities of the Profile Hidden Markov Models (pHMM). As a result, the central aspects of the underlying pHMM can be revealed through the visualization. Roca [25] proposed another profile-based approach to depict the frequency of residues at each position in combination with the traditional table approach. While a profile is an abstraction of the underlying MSAs that helps to reduce the amount of data that needs to be rendered, it does not capture important information such as the correlations between amino acid positions in subfamilies, and the visibility of specific subgroups in the MSAs. Clustering techniques are commonly used in both large-scale MSAs and microarray analysis visualization [26, 13], which shares many common challenges. Darling et al. [6] proposed a clustering approach to efficiently divide each sequence into blocks and achieve a less cluttered visualization. The authors also proposed matching criteria to assist the alignment process and highlight the relationships between sequences in the MSA. Clustering techniques can also be used to reduce the original MSA to a subset of representative sequences based on a sequence similarity metric. However, as the size of the original MSA increases, number of sequences in the clustering result still poses challenges to the traditional visual representation of MSAs.

Recent sequence analysis systems, such as Jalview [5, 33], and eBioX [1] do not only visualize MSA data, but also incorporate analysis capabilities. Users can perform annotations and alignment editing, such that domain knowledge can be incorporated to achieve a better alignment result. Jäger et al. [14] proposed a novel visual analytic approach to deal with time series alignment data. Their system allows users to align multiple series experiments and visually explore the aligned expression profiles through the integration of different 2D and 3D views. Miller et al. [20] have also proposed a multi-scale visualization approach that can handle dense evolutionary data. Moreover, other systems such as SEAVIEW and PHYLO_WIN allow users to easily modify parts of the alignment either manually or semi-automatically [7]. While current MSA visualization approaches and systems have shown to be useful in the analysis of MSA data, they all lack the ability of provide an overview of the whole set of sequences, from which gross trends can be identified. Recently, Herbig et al. [12] have proposed the GenomeRing visualization metaphor. By deriving a common coordinate system for all sequences in an MSA, visual annotations become possible. However, in contrast to our approach and the table-based layout adapted by most sequence renderers, GenomeRing exploits a circular layout. Although the circular layout nature of the technique helps to convey information efficiently for a small number of sequences, the visualization of large MSAs can be cluttered.

Gradient vector flow analysis. Active contours, or snakes, are extensively used in the field of computer vision and image processing to identify the boundaries of objects. Gradient Vector Flow (GVF), which serves as an external force field for active contours, is a feature-preserving spatial diffusion of gradients [34]. It overcomes the limitation of traditional active contour models proposed for instance by Kass et al. [15]. Particularly, GVF avoids the problem of getting stuck in boundary concavities and low capture range. Since its introduction, GVF has been successfully applied to dif-



(a) first 10 sequences limited to the first 117 columns



(b) all 71 sequences for all columns

Figure 2: Table-based visualization of the MSA from the Abp2 family (PF09441) using the Clustal color-coding scheme. While the visualization of the first 10 sequences limited to the first 117 columns in the MSA provides both, focus and context *(a)*, the visualization of the whole MSA makes the identification of features difficult *(b)*.

ferent applications of image processing. Ray and Acton applied the GVF to track leukocytes from intravital video microsopy [24]. While Bauer et al. [4] applied the GVF to extract skeletons from gray value images for virtual endoscopy, Hassouna et al. [11] applied the GVF to compute continuous curve skeletons from generic volumetric objects. Bauer et al. [3] also proposed a novel approach to use the GVF as a replacement for the multi-scale gradient computation in the application of tubular object detection in medical images. In this paper, we propose the use of GVF as a tool to convert pixel-based representations of large-scale MSAs into a flow of similarity and dissimilarity vectors, which enables us to detect global patterns, outliers, as well as data distributions within the MSAs. To our knowledge, this is the first approach exploiting GVF for MSA analysis and visualization.

3 Gradient Vector Flow Analysis

Figure 2 shows two table-based alignment representations of the Abp2 family (PF09441) containing 71 sequences using the Clustal color-coding scheme. Figure 2(a) manages to reveal both details down to the amino acid level and the feature patterns such as highly conserved areas, as well as non-conserved areas. However, the visualization of the whole 71 sequences in the MSA in Figure 2(b) shows already the limitation of this approach. As the number of sequences increases, the sole use of a color-coding scheme makes it more difficult to perceive gaps in columns. Consequently, it is more difficult to identify absolutely or highly conserved areas as well as areas containing no observable conservation.

In order to achieve a better overview visualization, we exploit the characteristics of patterns in large-scale MSAs. In highly conserved areas, there is a strong downward transition of the color patterns. On the other hand, the color patterns change disruptively or present sideway transitions in the areas containing gaps inserted by the alignment algorithms. In order to convey these directional patterns more clearly, we employ a vector field representation of the MSA data. Therefore, we exploit GVF analysis that enables feature extraction and highlighting of all, continuous structures, conserved areas, and the disruptive changes resulting from gap insertion of amino acid replacements during the alignment process.

3.1 Mathematical Background

Active contours are curves that move in an image while trying to minimize their energy. The traditional active contour models, or the snake algorithm, proposed by Kass et al. [15] uses an external force and an internal force to conform the contour to certain features in the image. While the external force is computed from the input image, the internal force is derived from the contour itself. The contour can be computed as follows. Let $s \in [0,1]$ be a parametric curve, the curve that represents the snake, $X_t(s,t)$, at time t is defined as follows:

$$X_t(s,t) = \alpha X''(s,t) - \beta X'''(s,t) - \nabla E_{\text{ext}} \tag{1}$$

where $X(s) = \big[x(s), y(s)\big]$, X'' and X''' are the second and fourth order derivatives, α and β are constants that define the internal forces, and ∇E_{ext} is the external force.

The drawback of this approach is that the algorithm often gets stuck in boundary concavities and is sensitive to its initialization. The GVF proposed by Xu and Prince [34] introduces a new external force to overcome the limitation of the traditional snake algorithm. The external force, ∇E_{ext}, in Equation 1 is replaced by a vector field, $V(x,y) = \big[u(x,y), v(x,y)\big]$, that minimizes the energy function:

$$E(V) = \int\int \mu(u_x^2 + u_y^2 + v_x^2 + v_y^2) + |\nabla f|^2 |V - \nabla f|^2 dxdy \tag{2}$$

where $f(x,y)$ is an edge map of the orignal image, ∇f is the corresponding gradient map, and μ is a constant that represents the noise level in the image. The solution of Equation 2 can be found by considering the vector field, $V(x,y)$, as a function over time:

$$\begin{array}{rcl} u_t(x,y,t) & = & \mu\nabla^2 u(x,y,t) - b(x,y)u(x,y,t) + c_1(x,y) \\ v_t(x,y,t) & = & \mu\nabla^2 v(x,y,t) - b(x,y)v(x,y,t) + c_2(x,y) \end{array} \tag{3}$$

where $b(x,y) = f_x(x,y)^2 - f_y(x,y)^2$, $c_1(x,y) = b(x,y)f_x(x,y)$, $c_2(x,y) = b(x,y)f_y(x,y)$, and ∇^2 is the Laplacian operator. Here, the iterative numerical solution of Equation 2 can be obtained by substituting the following variables to Equation 3.

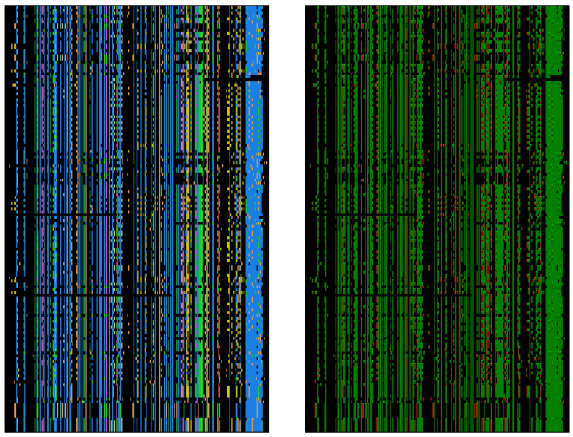
$$\begin{array}{rcl} u_t = \frac{1}{\Delta t}(u_{i,j}^{n+1} - u_{i,j}^n), \qquad v_t = \frac{1}{\Delta t}(v_{i,j}^{n+1} - v_{i,j}^n) \\ \nabla^2 u & = & \frac{1}{\Delta x \Delta y}(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}) \\ \nabla^2 v & = & \frac{1}{\Delta x \Delta y}(v_{i+1,j} + v_{i-1,j} + v_{i,j+1} + v_{i,j-1} - 4v_{i,j}) \end{array}$$

It is worth noting that the values for each position can be calculated independently of the other positions; thus, the GVF calculation for the whole MSA data can be parallelized. As a result, we can exploit the parallelism nature of the GPU to improve the performance of the algorithm.

3.2 MSA Analysis

The GVF algorithm discussed above contains two main stages: the computation of the initial GVF field and the propagation of the vector field based on the iterative solution of the underlying function. To enable the usage of GVF in the context of MSA analysis, we introduce constraints to both of these stages to directly incorporate the conserved structures as well as the disruptive changes into our algorithm.

Gradient initialization. We first represent the MSA as a pixel map, in which each line is a sequence in the MSA and each pixel is a residue of the corresponding sequence. The corresponding pixel color is represented in the HSV color space to be able to separate residues based on their hue. However, depending on the used color-coding scheme, the distribution in the hue channel may not be uniform. Therefore, we perform a histogram equalization on the hue



(a) Clustal color scheme (b) Initial vector field

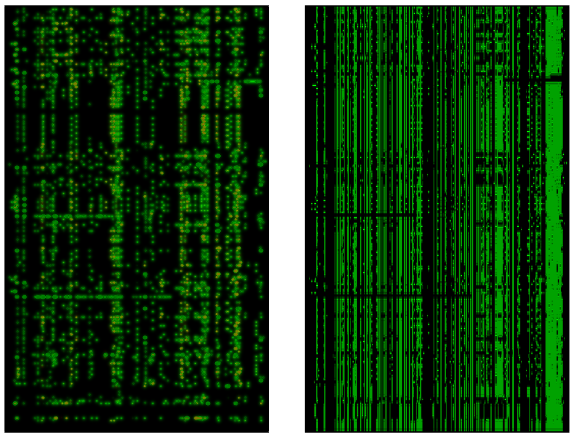
Figure 3: Vector field initialization based on the commonly used visual representation of MSAs. The visualization of the L1R_F9L family (PF02442) using the Clustal color scheme where black pixels denote gaps (a), can be transformed into the initial vector field used by the GVF algorithm (b). The x and y components of the vectors are encoded in the red and green channel; whereby high intensities are associated to positive values.

channel to avoid bias in the vector calculation process. Based on this representation, we can now compute the initial gradient field, whereby we exploit to different vector calculation methods. For the residues we calculate similarity vectors pointing into the direction of highest similarity, while for the gaps we exploit standard central differences gradient computation to obtain dissimilarity vectors. In the context of MSA analysis, the level of conservation is represented by the frequency of residues in the same column. This property can be represented as a flow pattern of highly conserved regions that is vertically directed. On the other hand, the disruptive change areas caused by the alignment process can be represented by a flow pattern that goes side ways. Therefore, the initial gradient field constructed from the input MSA image, I, is defined as follows:

$$grad(x,y) = \begin{cases} \big(0, 1.0 - (I(x,y+1) - I(x,y-1))\big) & \text{for residues} \\ \left(\frac{I(x+1,y)-I(x-1,y)}{2.0}, \frac{I(x,y+1)-I(x,y-1)}{2.0}\right) & \text{otherwise} \end{cases}$$

Figure 3(a) shows the visualization the L1R_F9L family (PF02442) using the Clustal coloring scheme where black pixels depict gaps in the alignment, and Figure 3(b) illustrates the initial vector field. To depict the vector field, we have color coded the x and y components in the red and green channel, and high intensities depict large positive values. While the original structure of the MSA is preserved, the constructed vector field captures the transition in both conserved and non-conserved areas.

Gradient propagation. In the second stage of the algorithm, the initial gradient field is propagated in such a way that original features of the input image are kept intact. In order to emphasize the features in the MSA, we introduce constraints into the propagation of the vector field. To achieve this, we make use of a weighting scheme applied to the Laplacian operator, ∇^2, in Equation 3. A uniform weighting scheme is used to depict outliers, which are emphasized through isotropic vector propagation into all directions. In addition, we also introduce the constraint to avoid the connection of different residues in the same column that leads to a false highly conserved area. In contrast, to emphasize the gaps introduced by the alignment algorithm in use, we constrain the propagation of the flow to a side-ways transition (see Figure 4(a)). Furthermore, constraining the propagation into the vertical direction allows us to extract the different levels of conservation in the MSA (see Figure 4(b)).

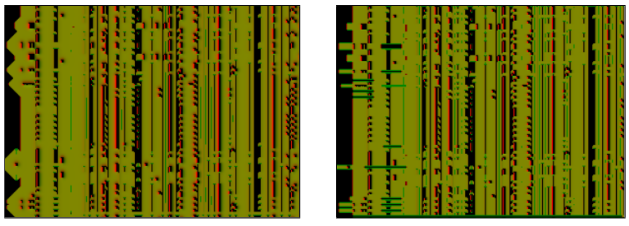


(a) Horizontal propagation (b) Vertical propagation

Figure 4: Result of the GVF analysis. Horizontal propagation enables the extraction of areas containing gaps as well as non-conserved residues (a), while vertical propagation allows us to identify highly or absolute conserved regions (b).

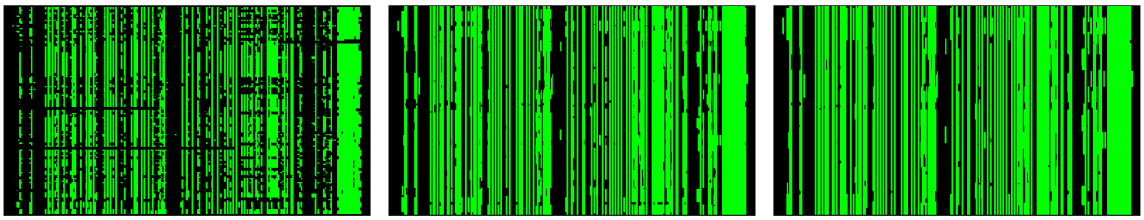
Figure 5 shows a comparison of isotropic and anisotropic propagation of the initial gradient field. In the standard isotropic propagation scheme, the initial gradient field is equally propagated in all directions as long as it minimizes the energy function in Equation 2. As can be seen in the top left of Figure 5(a), the isotropic property allows the gradients to propagate in all directions; thus, generate a circular shape (in shade of yellow) around the disruptive change (red). On the other hand, the initial gradient field is propagated unequally along different axes in the anisotropic propagation scheme (see Figure 5(b)). This enables us to extract highly conserved areas by emphasizing the propagation in the vertical direction. While the isotropic propagation scheme requires a smaller number of iterations to identify features of interest, it leads to the merging of regions that are close to each other, which is not always preferable for identifying conserved areas. However, the isotropic nature of the propagation process helps to identify the areas containing residues of low conservation. For instance, the isotropic propagation leads to the almost uniform distribution of gradient directions around the residues of low conservation. As a result, we can detect these features of interest through critical point detection.

As the gradient field propagation is an iterative process, the number of iteration can have impact on the result of the analysis. Figure 6 illustrates the results of anisotropic gradient field propagation in 0, 10, and 20 iterations. As the number of iterations increases, the gradient field propagation has the tendency to close the gaps between sequences in the MSA. Thus, we are able to extract regions,



(a) Isotropic propagation (b) Anisotropic propagation

Figure 5: Visualization of isotropic (a) and anisotropic (b) gradient propagation in the GVF analysis. While isotropic propagation emphasizes residues of low conservation, anisotropic propagation emphasizes highly conserved regions and gaps.



(a) 0 iterations (initial vector field) (b) 10 iterations (c) 20 iterations

Figure 6: Results of the anisotropic vertical gradient field propagation with a different number of iterations: 0 iterations (a), 10 iterations (b), and 20 iterations (c). A higher number of iterations leads to a gap closure between same residues, and thus allows to detect areas with a high frequency of residues but no clear conservation.

where no clean conservation is present, but a high frequency of the same residues with only local interruptions. The analysis on this result can help to provide a clean overview of the MSA as demonstrated in Section 4.

4 Visualization

The advantage of the commonly used color-coded table-based representation is the ability to reveal spatial correlations as well as local structures within MSA data sets. However, the visualization becomes less expressive as the number of sequences in the MSA increases. To deal with this downside when visualizing large-scale MSA data, often the spatial context is neglected and the sequences are depicted through statistics, as for instance residues histograms. While profile-based visualizations [19, 27, 28] enable the abstraction of MSA by reducing the amount of information to be rendered, a profile does not capture any spatial information, such as the correlations between alignment positions, or the visibility of discernible sub-groups in the MSA. Consequently, such a visualization does not enable to obtain a detailed view of the whole MSA with all sub-structures. When on the other hand using sequence clusters to reduce the amount of data to be visualized, the spatial context can be preserved. However, large MSAs that contain tens of thousands of sequences still result in high a number of clustered sequences which is challenging for the table-based visualization approach. By exploiting the proposed GVF analysis, we are able to support the visual identification of patterns in large-scale MSA data. It is worth mentioned that the proposed GVF analysis is shape-preserving. Particularly, the proposed technique does not introduce artifact into the original input data. As a result, the extracted features from the GVF analysis conform with the original structure in the MSA under investigation, and the proposed visualization technique does not introduce artifact into the original data.

Feature emphasis. Based on the GVF analysis, we can identify conservation areas, the patterns of the gaps inserted by the alignment algorithm, as well as the areas containing residues of low conservation. The colors of these features have been chosen to comply with the rules of pre-attentive perception [31]. In the GVF, the flows that represent the conserved areas move vertically so they are colored green. On the other hand, the flows that moves horizontally, which reflect the gaps inserted by the alignment algorithm, cause disruptive changes so they are colored red. The residues with low conservations are colored in shades of yellow for highlighting. Figure 1(b) illustrates the visualization of features extracted from the GVF analysis applied to the L1R_F9L family (PF02442). While the green vertical lines depict the conserved areas, the red horizontal lines show the disruptive changes in the flow, which reflect the patterns of the gaps inserted by the alignment algorithm.

A common problem arising when visualizing large-scale MSA data is that the size of features can be small in comparison to the large amount of data that needs to be rendered. Therefore, in order

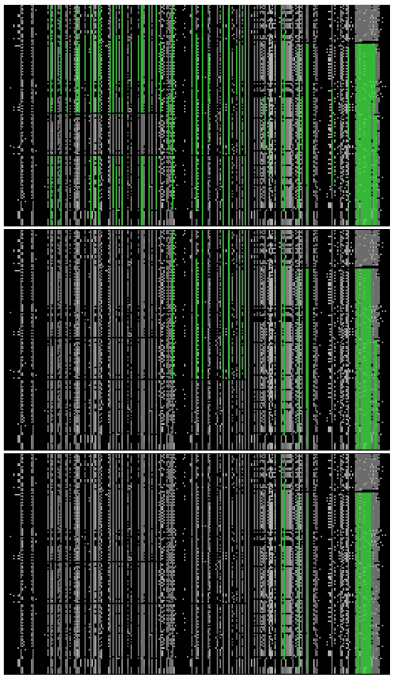
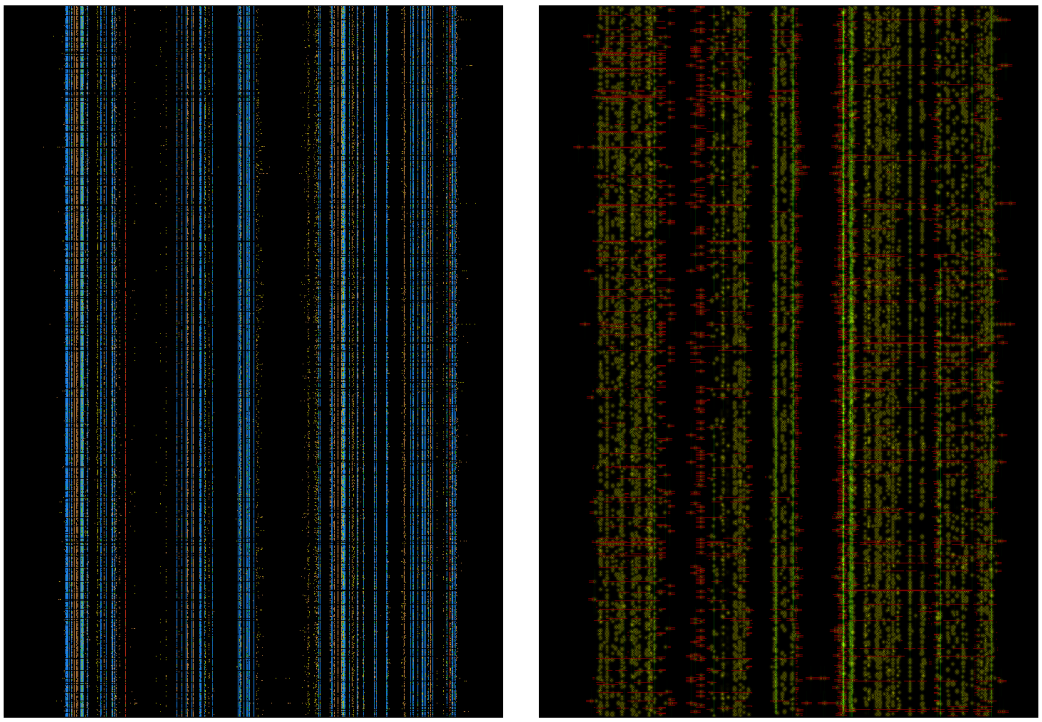


Figure 7: Level of conservation exploration. From top to bottom is the visualization of conserved areas that have the length greater than or equal to 20%, 50%, and 80% the length of the absolute conserved area.

to emphasize small features of interest in an overview visualization, special considerations are required. Particularly, while the widely used morphology operators [29] in image processing can help to highlight small features, they introduce artifacts along the way since they are not shape-preserving. The advantage of using the presented GVF analysis allows us to increase the propagation process of the gradient field to upscale the detected features and guarantee the coherence with the original structure of the MSA. In Figure 1(b), the residues with low conservation are shown as yellow-shaded circles and they are adhere to the original structure as shown in Figure 1(a).

Now that we were able to depict features of interest, they need to be shown in the context of the original representation in order to link them to specific residues. While multiple linked view setups have been proven to be very useful when dealing with multiple visual representations [32], its application to the visualization of large-scale MSAs can be limited due to the amount of rendered information as well as the complexity of underlying data distributions. As especially the latter makes the mental registration difficult, alternative concepts are needed. We make use of overlaying



(a) Clustal color scheme (b) GVF-based MSA features

Figure 8: Visualization of the MSA of the Sulfite exporter TauE/SafE (PF01925) consisting of 12,351 sequences with 1,220 residues using the Clustal color scheme (a), and the proposed visualization technique (b). Through GVF analysis, the disruptive changes and the level of conservation are captured and emphasized.

and use the extracted features as visual annotations. As illustrated in Figure 1(c), the detected features are color-coded and imposed onto the gray scale format of original MSA image, which serves as the background formation. This enables us to highlight features in the context of the original MSA and avoid a cumbersome mental spatial registration between the two visual representations. In addition, we allow for interactive changes of the degree of blending between the two representations.

Conservation exploration. While the standard conservation histogram provides statistical information about the distribution of the residues in the individual columns, it does not capture their spatial distribution. By using GVF propagation with the vertical constraint introduced above, we can identify the length of the flow patterns to detect conserved regions. As the detected features adhere to the original structures, we can not only show the level of conservation but also visualize them in the original context. This enables us to facilitate the exploration of the conservation level in the MSA. Figure 7 illustrates the exploration of different levels of conservation in the MSA. From top to bottom is the visualization of the conserved areas that have the length greater than or equal to 20%, 50%, and 80% of the absolute conservation. As the level of conservation exploration can help researcher to quickly identify absolute conserved areas as well as highly conserved areas, it facilitates the detection of residues that are critical for function as well as residues that are probably important for function or structure stability.

5 Test Cases

We applied the proposed visualization technique to two large-scale MSAs from the Pfam database (Wellcome Trust-Sanger Institute) [23]. The Sulfatases are a highly conserved enzyme family, with similar overall folds, mechanisms of action, and bivalent metal ion-bindings [8, 30]. The MSA (PF00884) contains 18,763 sequences with the maximum length of 1,668 residues, whereby the MSA of the Sulfite exporter TauE/SafE (PF01925) contains 12,351 sequences with the maximum length of 1,220 residues. In both cases, as the number of sequences exceeds ten thousand, the ability to convey the gross trends of the whole family is a challenge.

Figure 9 shows the visualization of the Sulfatase enzymes family (PF00884) using our proposed visualization approach. In Figure 9(a), the whole family is visualized using the Clustal color scheme. Each line in the image represents a sequence and each residue is represented as a pixel at the corresponding position. Figure 9(b) illustrates the result of the extracted features based on the GVF analysis. Although the use of color helps to reveal the highly conserved position in the traditional representation, it is difficult to perceive the patterns of the gaps that disrupt the conservation. Our visualization in Figure 9(b) not only manages to reveal the highly conserved regions but also to emphasize the patterns resulting from gaps. The visualization depicts highly conserved areas in a manually annotated data set, which has been previously visualized by Procter et al. [22]. With our approach, the areas containing residues of low conservation are detected through the GVF analysis and highlighted through yellow-shaded circular shapes. The application to the Sulfite exporter TauE/SafE as shown in Figure 8 has similar qualities. Thus, our visualization enables a quick overview by depicting the gross trends which are of potential interest. By seamlessly blending between the standard Clustal representation and our extracted features, domain experts can explore large-scale MSA data sets by obtaining an overview and residue details simultaneously.

6 Conclusions and Discussions

In this paper, we have proposed a novel approach for extracting and visually emphasizing patterns in MSA data. By transforming

a pixel-based MSA representation into a vector field, we were able to apply GVF analysis in order to extract patterns, which are important during the MSA analysis process. Therefore, we combine isotropic and anisotropic vector propagation, which helps us to extract conserved as well as non conserved areas. Due to the shape preserving nature of the GVF analysis, the extracted information can be used to visually emphasize the detected features, and ultimately annotate a standard table-based MSA representation. The proposed visual emphasis has been developed with respect to the perceptual properties of the human visual system, and thus enables the user to spot patterns of interest even in overview visualizations of large-scale MSA data. We have demonstrated our approach by applying it to various MSA data sets used for protein analysis.

In the future, we would like to conduct a thorough evaluation to serve as a foundation for improving the presented approach. Thus, extending the technique to handle genome data, and making it available to a wider audience of domain experts are important steps. While we currently derive the initial vector field from a histogram-equalized pixel-based MSA representation, we plan to experiment with different value distributions. By distributing the intensities encoding residues in various ways, it would become possible to specifically emphasize transitions between certain amino acids or amino acid groups of interest. Furthermore, we would like to investigate how the presented approach can be used to compare the results of different alignment algorithms, and how we can exploit different sequence orderings, which we currently not address.

Acknowledgements

This work was supported through grants from the Excellence Center at Linköping and Lund in Information Technology (ELLIIT), the Swedish Research Council (VR, grant 2011-4113), and the Swedish e-Science Research Centre (SeRC). The presented technique has been integrated into the Voreen volume rendering engine (`www.voreen.org`), which is an open source visualization framework.

References

[1] Á. Barrio, E. Lagercrantz, G. Sperber, J. Blomberg, and E. Bongcam-Rudloff. Annotation and visualization of endogenous retroviral sequences using the Distributed Annotation System (DAS) and eBioX. *BMC bioinformatics*, 10(Suppl 6):S18, 2009.

[2] G. Barton et al. ALSCRIPT: a tool to format multiple sequence alignments. *Protein Engineering Design and Selection*, 6(1):37–40, 1993.

[3] C. Bauer and H. Bischof. A Novel Approach for Detection of Tubular Objects and Its Application to Medical Image Analysis. In *Pattern Recognition*, pages 163–172. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[4] C. Bauer and H. Bischof. Extracting Curve Skeletons from Gray Value Images for Virtual Endoscopy. In *Medical Imaging and Augmented Reality*, pages 393–402. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[5] M. Clamp, J. Cuff, S. Searle, and G. Barton. The Jalview java alignment editor. *Bioinformatics*, 20(3):426–427, 2004.

[6] A. Darling, B. Mau, F. Blattner, and N. Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, 14(7):1394–1403, 2004.

[7] N. Galtier, M. Gouy, and C. Gautier. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Computer applications in the biosciences: CABIOS*, 12(6):543–548, 1996.

[8] D. Ghosh. Human sulfatases: A structural perspective to catalysis. *Cellular and Molecular Life Sciences*, 64(15):2013–2022, June 2007.

[9] L. Goodstadt and C. Ponting. CHROMA: consensus-based colouring of multiple alignments for publication. *Bioinformatics*, 17(9):845–846, 2001.

[10] P. Gouet, E. Courcelle, D. Stuart, et al. ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics*, 15(4):305–308, 1999.

[11] M. S. Hassouna and A. A. Farag. On the Extraction of Curve Skeletons using Gradient Vector Flow. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, 2007.

[12] A. Herbig, G. Jger, F. Battke, and K. Nieselt. Genomering: alignment visualization based on supergenome coordinates. *Bioinformatics*, 28(12):i7–i15, 2012.

[13] M. A. Hibbs, N. C. Dirksen, K. Li, and O. G. Troyanskaya. Visualization methods for statistical analysis of microarray clusters. *BMC bioinformatics*, 6(1):115, 2005.

[14] G. Jager, F. Battke, and K. Nieselt. TIALATime series alignment analysis. In *Biological Data Visualization (BioVis), 2011 IEEE Symposium on*, pages 55–61. IEEE, 2011.

[15] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[16] G. Landan and D. Graur. Characterization of pairwise and multiple sequence alignment errors. *Gene*, 441(1-2):141–147, 2009.

[17] M. Larkin, G. Blackshields, N. Brown, R. Chenna, P. McGettigan, H. McWilliam, F. Valentin, I. Wallace, A. Wilm, R. Lopez, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007.

[18] K. Lin, A. May, and W. Taylor. Amino acid encoding schemes from protein structure alignments: Multi-dimensional vectors to describe residue types. *Journal of theoretical biology*, 216(3):361–365, 2002.

[19] M. Madera, C. Vogel, S. K. Kummerfeld, C. Chothia, and J. Gough. The SUPERFAMILY database in 2004: additions and improvements. *Nucleic acids research*, 32(suppl 1):D235–D239, 2004.

[20] R. Miller, V. Mozhayskiy, L. Tagkopoulos, and K. Ma. EVEVis: A multi-scale visualization system for dense evolutionary data. In *Biological Data Visualization (BioVis), 2011 IEEE Symposium on*, pages 143–150. IEEE, 2011.

[21] D. Morrison. Sequence alignment: Methods, models, concepts, and strategies. *Systematic Biology*, 59(3):363–365, 2010.

[22] J. Procter, J. Thompson, I. Letunic, C. Creevey, F. Jossinet, and G. Barton. Visualization of multiple alignments, phylogenies and gene family evolution. *Nature methods*, 7:S16–S25, 2010.

[23] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn. The Pfam protein families database. *Nucleic acids research*, 40(D1):D290–D301, Dec. 2011.

[24] N. Ray and S. T. Acton. Motion gradient vector flow: an external force for tracking rolling leukocytes with shape and size constrained active contours. *Medical Imaging, IEEE Transactions on*, 23(12):1466–1478, 2004.

[25] A. I. Roca, A. E. Almada, and A. C. Abajian. ProfileGrids as a new visual representation of large multiple sequence alignments: a case study of the RecA protein family. *BMC bioinformatics*, 9(1):554, 2008.

[26] P. Saraiya, C. North, and K. Duca. An evaluation of microarray visualization tools for biological insight. pages 1–8, 2004.

[27] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, 18(20):6097–6100, 1990.

[28] B. Schuster-Böckler, J. Schultz, and S. Rahmann. HMM Logos for visualization of protein families. *BMC bioinformatics*, 5(1):7, 2004.

[29] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, Inc., Orlando, FL, USA, 1983.

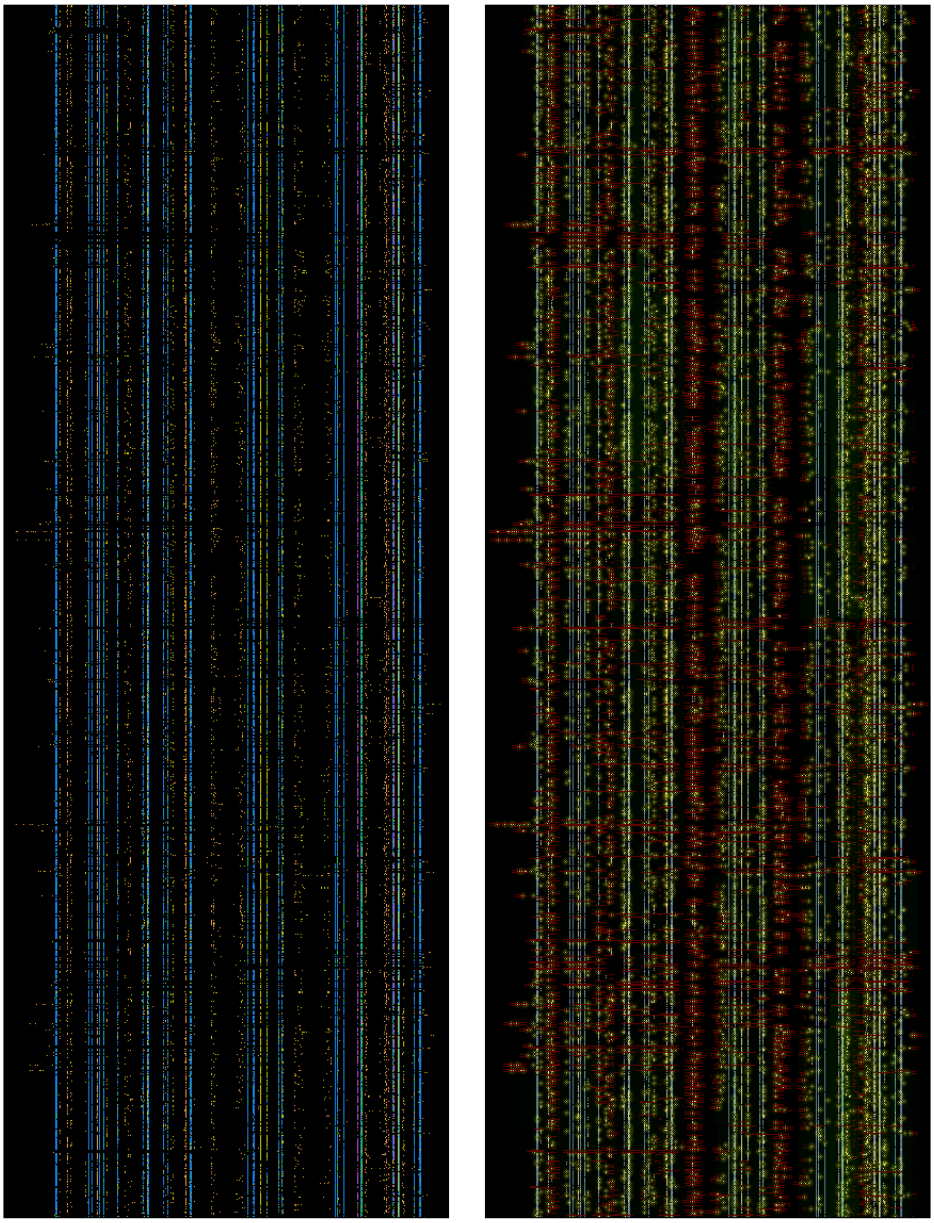
[30] The UniProt Consortium. The Universal Protein Resource (UniProt) 2009. *Nucleic acids research*, 37(Database):D169–D174, Jan. 2009.

[31] J. Theeuwes. Perceptual selectivity for color and form. *Perception & Psychophysics*, 51(6):599–606, 1992.

[32] M. Tory. Mental registration of 2D and 3D visualizations (an empirical study). In *IEEE Visualization*, pages 371–378, 2003.

[33] A. Waterhouse, J. Procter, D. Martin, M. Clamp, and G. Barton. Jalview version 2a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, 2009.

[34] C. Xu and J. L. Prince. Snakes, shapes, and gradient vector flow. *Image Processing, IEEE Transactions on*, 7(3):359–369, 1998.



(a) Clustal color scheme (b) GVF-based MSA features

Figure 9: Visualization of the MSA of Sulfatase (PF00884) consisting of 18,763 sequences with the longest length of 1668 residues. The visualization of the whole MSA using the Clustal color scheme (a), in comparison to the features extracted and visualized using the proposed GVF analysis technique (b). While the vertical green lines depict the highly conserved areas, the horizontal red lines show the patterns of the gaps inserted by the alignment algorithms that cause disruption in the conservation areas. The residues of low conservation are shown in yellow-shaded circular shapes. The proposed technique enables highlighting of the absolutely or highly conserved areas as well as emphasizing the gaps that cause disruption.