Mu-8: Visualizing Differences between a Protein and its Family

John Mercer[*]
Harvard University
Broad Institute

Balaji Pandian [†]
Harvard University

Nicolas Bonneel [‡]
Harvard University

Alexander Lex [§]
Harvard University

Hanspeter Pfister [¶]
Harvard University

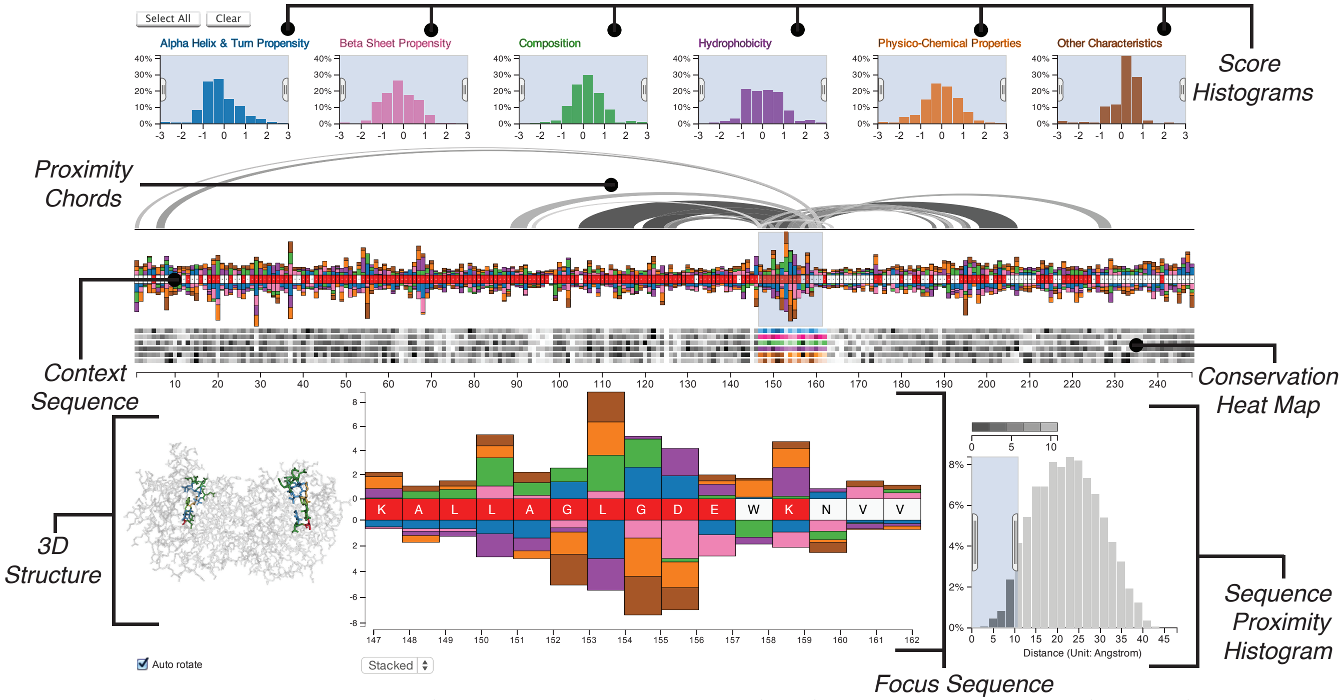

Figure 1: The annotated Mu-8 interface showing how characteristics of a defective protein compare to its functional family.

Abstract

A complete understanding of the relationship between the amino acid sequence and resulting protein function remains an open problem in the biophysical sciences. Current approaches often rely on diagnosing functionally relevant mutations by determining whether an amino acid frequently occurs at a specific position within the protein family. These methods, however, fail to appropriately account for the biophysical properties and the 3D structure of the protein. To remedy this, we have developed an interactive visualization technique, *Mu-8*, that provides researchers with a holistic view of the residues that have significantly different characteristics from a family of homologous yet functional proteins. Mu-8 enables analysts to identify regions of the sequence that have biophysical anomalies, while clearly communicating the spatial relationships amongst residues.

Index Terms: Protein Function, Genetic Variants, Amino Acid Indices, Biological Visualization.

1 Introduction

Proteins are biochemical products that perform specific functions in a cell or an organism. A protein is made of a sequence of amino acids (also referred to as residues) that are coded for by genes. Proteins perform vital roles, including metabolic processes and housekeeping, such as DNA replication. A protein derives its function from its three-dimensional structure (the tertiary structure), which is in turn driven by the biochemical properties of its amino acid sequence (the primary structure). Understanding and being able to predict the 3D structure from the amino acid sequence, however, is part of the unsolved protein-folding problem [4].

While a general solution to this problem is not within reach of current methods, interactive visualization and computational analysis can help biologists understand the relationship between the amino acid sequence and a protein's 3D structure. This in turn will enable analysts to predict which mutations in an amino acid sequence cause the loss of function in a protein.

Motivated by the problem and the data published for the 2013 IEEE BioVis Data Contest[1], we have developed *Mu-8*, a novel, interactive visualization tool for comparing proteins to their family and for identifying potential regions that cause functional breakdown. Different or altered proteins can often fulfill the same function, albeit often with different efficiency. Such proteins are referred to as a protein family and are mostly evolutionary related. This demonstrates that function is often preserved even if the amino acid sequence is changed. On the other hand, small changes to the sequence can sometimes cause function to break down. The challenge posed by the BioVis Data Contest is to find out which mutation(s) in a highly mutated amino acid sequence causes this functional break-down. Using Mu-8, an analyst can: (1) quickly identify residues or regions of residues that are significantly different from the family with respect to one or more characteristics; (2) identify whether such a region is in an otherwise highly conserved

[*] e-mail: mercer@broadinstitute.org
[†] e-mail: balajipandian@college.harvard.edu
[‡] e-mail: nbonneel@seas.harvard.edu
[§] e-mail: alex@seas.harvard.edu
[¶] e-mail: pfister@seas.harvard.edu

[1] `http://biovis.net/year/2013/info/contest`

area of the sequence; and (3) assess the spatial relationships to other regions of the sequence.

Mu-8 was able to identify several regions of interest. Most notable are the residues at positions 150-156, which mutated from "VLEEVKD" to "LAGLGDE" (see Fig. 1 in the focus region). These residues are significantly different from the family across many biophysical properties, are located in relatively conserved regions, and are close to other regions with similar anomalies in the folded protein. This region is also close to the protein's active site as lysine 12, histidine 95, and glutamic acid 165 are directly involved in the metabolic process [6]. It stands to reason that the mutated region 150-156 may have contorted the location and orientation of the active site, thus rendering the protein dysfunctional.

2 Concept

Our design strategy was predicated on a few basic principles, which we elicited in interviews with domain experts and an extensive literature review. First, we required a design that captured the most important information in the vast number of amino acid indices (which we describe herein). Second, we required both a holistic view of the quantitative landscape of the sequence and the ability to focus in on regions of interest. Finally, we wanted to unite the analysis of the sequence with the inspection of the 3D structure.

To this end, for each of the amino acid's six major characteristics, we developed a score - which we call *c-scores* - to measure how different a residue in the mutated protein is compared to the larger protein family. These scores are derived from amino acid indices, which are an invaluable resource for judging the potential consequence of a mutation. An amino acid index is a quantitative score assigned to each of the amino acids. They predict various biophysical properties and their development has become a mainstay in protein research pioneered by Chou and Fasman [3].

However, there are hundreds of amino acid indices, and determining which of them are relevant to the loss of function is difficult. At the same time, showing all indices in a visualization is a challenge with respect to scalability and introduces significant complexity. To address this problem we employ principal component analysis on the indices within the six major characteristics, thereby isolating the most important information while reducing complexity. We then use the first principal component as the main ingredient to the c-scores.

Amino acids substitute for one another with varying degrees of efficiency. Our approach is based on the idea that significantly different characteristics of substituted amino acids are more likely to cause functional changes. Consequently, our c-scores are adjusted to magnify how "different" a substitute amino acid is from its equivalent in the functional protein family. Mutations affecting function often occur in otherwise conserved regions, which are regions with low variation of residues in homologous proteins. Our scores also account for this variation in the family. The distribution of these c-scores are shown in the *Score Histograms*, while the individual scores for each amino acid are shown as bars in the *Context Sequence* and *Focus Sequence* views (see Figure 1). To complement these scores we also highlight conserved regions with a *Conservation Heat Map*, also shown in Figure 1, which shows how conserved an amino acid is with respect to each characteristic.

A recurring theme in our research has been the paramount importance of the spatial context of an amino acid. We address this by incorporating 3D structural information into the visualization in two ways: (1) we use chords to connect the residues within a specified distance of a selected group of residues (thus identifying the "sphere of influence" of a region of the sequence); and (2) we include a 3D rendering of the functional protein.

3 Data

While our approach is easily generalized, we demonstrate Mu-8 using the defective *triose-phosphate isomerase (TIM)* sequence published as part of the BioVis Contest. TIM enzymes are utilized in glycolysis, an important metabolic process, and are essential for energy production. The enzyme is found in all living organisms and, in the case of humans, mutations can cause a severe metabolic disease called *triosephosphate isomerase deficiency*. We were provided with a functioning TIM isolated from *Saccharomyces cerevisiae (scTIM)* [6], a family of functional TIMs, and a defective TIM (dTIM) created from mutating scTIM [7].

We acquired data of the protein sequence, three-dimensional positions [7], and a set of amino acid indices from the *GenomeNet AAindex database* [5, 9]. In this section, we detail our processing of this information.

3.1 Sequence Data

The sequence data includes dTIM (non-functional), scTIM (functional parent of dTIM), and a set of 5,508 other TIMs which we call the *family*. The length of both dTIM and scTIM is 248 residues, while other TIMs vary between 23 and 1053 with an average of 228 residues. To incorporate TIMs of different lengths, we conducted a multiple sequence alignment using the *Clustal* software [2]. After aligning the sequence data, we cropped off the amino acids outside of the 248-residue window of dTIM and scTIM.

3.2 Index Data

Amino acid indices are quantitative measures of molecular characteristics. We study indices pertaining to six characteristics, for a total of 442 indices, originally analyzed by Tomii and Kanehisa [9]. These include:

- **alpha and turn propensity**, which quantifies the likelihood of forming an α-helix (118 indices),
- **beta propensity**, which quantifies the likelihood of forming a β-sheet (37 indices),
- **hydrophobicity**, which quantifies how water-repellent an amino acid is (149 indices),
- **composition**, which quantifies the types of atoms that comprise each amino acid (24 indices),
- **physicochemical properties**, which quantifies physical and chemical characteristics such as bulkiness (46 indices), and
- **other properties**, which describes indices that do not fit within the other 5 categories, such as the likelihood that an amino acid will be located on the surface of the protein (28 indices).

An example index from the alpha and turn propensity group, developed by Prabhakaran [8], provides a score for the relative frequency of a residue in an alpha-helix structure, and is defined as the ratio of the observed to expected frequency of the residue in the alpha helix structure. Residues with greater than expected observed frequency have an index greater than one.

There is of course variation amongst, and uniqueness to, these hundreds of indices. However, our design objective was to provide a holistic view of each characteristic across the entire sequence. Therefore, we reduce the dimensionality of the problem using the method of principal components. We found that the first principal component accounts for a significant proportion of variability (between 50% and 75% for the 6 characteristics for the TIM data) and therefore chose it as the central metric for our scores. For each of the six characteristics c, we perform the following procedure to compute *characteristics scores (c-scores)*. First, we calculate the first principal component of the indices for each characteristic of the dTIM (1 protein x 248 residues) and the aligned sequences of the family (5509 proteins x 248 residues). For the family, we also compute the average and the standard deviation of the principal component at each position. Finally, we use this information to calculate the characteristics score $cs_{dTIM}^{c,r}$ for each residue using the formula

$$cs^{c,r}_{dTIM} = \frac{PC1^{c,r}_{dTIM} - \overline{PC1^{c,r}_{fam}}}{\sigma_{PC1^{c,r}_{fam}}}, \quad (1)$$

where $PC1^{c,r}_{dTIM}$ denotes the first principal component of characteristic c for residue r of the dTIM sequence, $\overline{PC1^{c,r}_{fam}}$ denotes the average first principal component of characteristic c for residue r across the family, and $\sigma_{PC1^{c,r}_{fam}}$ is the standard deviation of the first principal component of the family for characteristic c of residue r.

The impetus for this metric is to identify locations of the sequence such that the first principal component of dTIM is significantly different from the family mean in positions that are highly conserved. Significantly high or low scores at points of mutation highlight residues of dTIM that may be culprits in function loss.

3.3 3D Structure and Proximity Data

We use the three-dimensional PDB model of scTIM [6] for our 3D view. Pairwise distances between the α-carbons of each amino acid are computed to determine whether two amino acids are within each other's sphere of influence.

4 The Mu-8 Interface

In this section we discuss the design rationale for the visual encodings of the sequence, the c-scores, our measure of conservation, the 3D structure and the proximity data, which, in concert provide the analyst with the desired holistic view.

4.1 Score Histograms

The six histograms at the top of the visualization (see Figure 1) show the distributions of the c-scores, conveying dTIM's difference to the family across the entire sequence. The tails of these distributions encode for residues that have either a significantly greater or smaller c-score than the family, i.e., the amino acids at the tails behave significantly different than the family. The histograms use a uniform y-axis and are capped at ±3 standard deviations to counter-balance the effects of outliers. The histograms can be used to filter scores in a selected range. Figure 2, for example, shows a filter excluding all scores outside the -2 to -0.5 interval. This is especially useful to select the tails of the distribution to highlight, for example, all amino acids that have a strongly increased hydrophobicity compared to the family consensus. Each histogram is given a unique color to identify the characteristics, which corresponds to the color of the bars in the sequence views. Regions of the histogram that are filtered-out are shown in gray.

4.2 Sequence Views

At the center of Mu-8 are two sequence views which are used to encode the mutation status, the c-scores and the degree of conservation of the residues. The *context sequence view* shows the whole sequence of amino acids from left to right. Red residues indicate that these are mutated with respect to the parent protein while white residues indicate consensus with the parent protein. A labeled axis below the sequence facilitates orientation and enables analysts to easily reference regions.

Above and below the sequence we show stacked bars showing the c-scores for each characteristic thus highlighting the cumulative deviation from the family. Characteristics with a positive c-score are stacked on top of the sequence, while those with a negative score are stacked below the sequence. Figure 2 shows an example for the relationship of the histograms to the amino acid sequence. For the part of the sequence shown, two amino acids have scores matching the filter specified in the histogram, thus the corresponding bars are rendered.

While the context sequence view provides a convenient overview of the whole sequence, details such as the specific amino acid or the exact scores are obscured. To remedy this we supplement the context sequence view with a *focus sequence view* (see Figure 1), which provides a larger version of a selected region of interest. The selected region is specified using a window on the context region, the size of which can be dynamically adjusted, but is limited to 15 residues to ensure readability of the focus sequence.

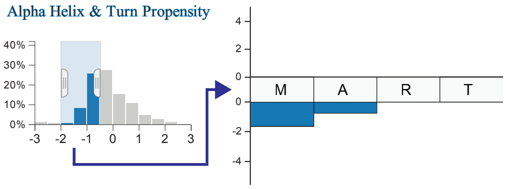

Figure 2: Filtering of c-scores between -0.5 and -2 for the *alpha helix & turn propensity* characteristic. An example of how such a score is mapped to the sequence is shown on the right.

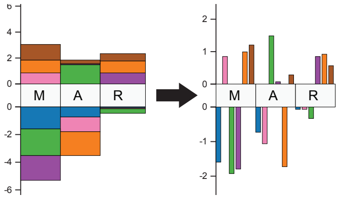

Figure 3: Stacked bars compared to aligned bars for several residues.

The stacked bars used in the context sequence allow an analyst to easily judge the overall deviation from the family. Judging the magnitude of the individual scores, however, is difficult using the stacked bars, as relative lengths of not-aligned elements are perceptually more difficult to distinguish compared to judging relative lengths of aligned elements. In the focus view, we provide the option to switch c-scores from a stacked to an aligned bar chart - which facilitates detailed comparisons within and amongst residues.

4.3 Conservation Heat Map

Below the context sequence view is the *conservation heat map*, also shown in Figure 1. For each characteristic, this heat map encodes the variation of residues in the family. Conserved regions are known to be more relevant for function, since evolutionary pressure selects for functional proteins, while variable regions often are less relevant for function. Conservation is considered when calculating the c-scores, which results in higher scores for deviations in highly conserved regions. The additional heat map enables the analyst to (a) judge conservation independently of effect size and (b) judge the relevance of outliers. In the heat map dark cells encode a high variability, while bright cells encode for a conserved residue. Which line corresponds to which characteristic can be identified through the color used in the selected window. We use an HSL color scale to match the perceived brightness of the gray-scale and the colored areas.

4.4 Visualizing Proximity

Changes in the biochemical properties of the sequence influence the folding and thus the function of a protein. A linear representation of the amino acid sequence, as used in the sequence views, however, cannot adequately account for the biochemical spheres of influence of the residues. Therefore we supplement the sequence view with proximity chords and provide a 3D structure view.

The *proximity chords* connect the focus region of the sequence with other residues that are within a user-specified distance from the focus region, as shown on top of the context sequence view in Figure 1. The sphere of influence that is of interest is dependent on the type of analysis. To account for this we provide the analyst with the means to specify the proximity using the *sequence proximity*

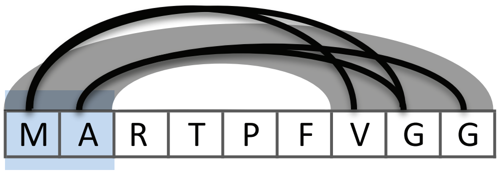

Figure 4: The residues in the focus region are within the specified distance of three adjacent residues further down the sequence, as illustrated by the black arcs. To reduce visual clutter, we replace the arcs connecting individual residues with chords (shown in gray) connecting proximate regions.

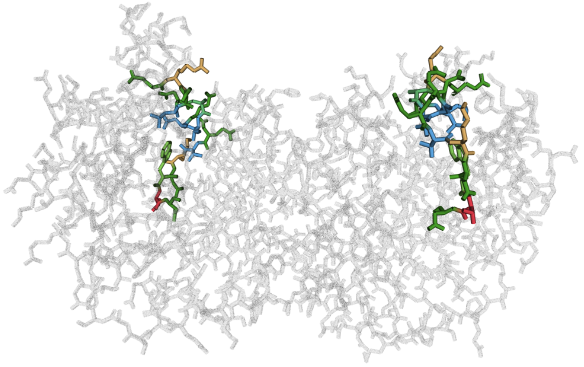

Figure 5: Complementary 3D structure linked to the sequence. Selected residues are color coded.

histogram, shown at the lower right of Figure 1. This histogram shows the distribution of the distances of all residues relative to the residues in the focus region. By brushing the histogram, the analyst can specify the relevant proximity, which in turn filters the chords above the sequence. The chords are rendered at varying brightnesses, with darker chords encoding closer residues and brighter chords encoding more distant residues, as encoded in the legend above the histogram.

Due to the nature of the sequence, it is natural that the immediate neighborhoods of a residue are at similar distances to other neighborhoods in the sequence. We use this observation to reduce the visual clutter of the chords by bundling regions with similar proximity, as illustrated in Figure 4. In this example, the two residues in the focus region (M and A) are all connected to three residues adjacent to each other (V, G, and G). Instead of rendering a chord for every residue, as shown in black, we bundle them to a wider arch, as shown in gray.

4.5 Visualizing 3D Structure

The 3D structure is critical for the function of the protein. We account for this with an all-atom visualization (omitting hydrogen atoms) of the three-dimensional structure of the functional protein, scTIM, as shown at the bottom left of Figure 1 and in detail in Figure 5. The residue centric paradigm of Mu-8 motivated the all-atom representation over a visualization of the secondary structure or the protein surface.

The view can be rotated, zoomed, and panned to inspect neighborhoods more closely. It is also linked to the sequence views such that the residues in the focus region are highlighted using an established color scheme for amino acids [1].

5 Implementation and Scalability

We pre-processed the data using R and C code. The visualization uses the D3 JavaScript library, with the exception of the 3D view, which employs WebGL. Mu-8 is open source, the code and data are accessible through the project website[2]. We tested our implementation on recent versions of Google Chrome and Mozilla Firefox. Microsoft Internet Explorer currently does not support WebGL and thus can not be used to run Mu-8.

Mu-8 scales well for the needs of the dTIM protein and its family. We expect Mu-8 to also handle larger proteins, but estimate an upper limit of approximately 1000 amino acids, as at this point an amino acid would be represented by less than two pixels on a full-HD screen. While this makes Mu-8 applicable to the majority of all proteins, there are some that exceed this size considerably, which would require a modified approach.

6 Conclusion and Future Work

We contend that Mu-8 is a comprehensive visual analysis solution to compare differences between a protein and its family. Our approach elucidates the significant biochemical differences while accounting for conservation, proximity amongst residues, and overall 3D structure. As previously mentioned, Mu-8 reveals several candidate regions that may cause function to break down in the dTIM protein. The most notable mutated region is "LAGLGDE" located at positions 150-156. The evidence suggests that this region is: (1) significantly different across several characteristics; (2) is relatively conserved; (3) close to other regions that exhibit suspect behavior in the folded protein; (4) close to the proteins active site.

In the future, we intend to generalize the Mu-8 approach so that researchers can employ Mu-8 for other protein data. To this end, we intend to provide a web-service that processes the input data and produces the Mu-8 interface.

We also consider improving our scores by including more principal components to account for more of the variation across the indices. In addition, a search and selection feature to enable the analyst to select a specific amino acid index (for a given characteristic) in addition to or instead of the c-scores would be desirable.

Acknowledgements

The authors wish to thank the BioVis contest team for helpful clarifications and Dr. Guido Guidotti, the Higgins Professor of Biochemistry at Harvard University, for his helpful feedback. This work was supported in part by the Austrian Science Fund (J 3437-N15), the National Science Foundation (CGV-1111415) and the Air Force Research Laboratory and DARPA grant FA8750-12-C-0300.

References

[1] B. Bodenmiller. Amino acid colour schemes, Apr. 2006.

[2] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. Multiple sequence alignment with the clustal series of programs. *Nucleic acids research*, 31(13):3497–3500, July 2003.

[3] P. Y. Chou and G. D. Fasman. Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in enzymology and related areas of molecular biology*, 47:45–148, 1978. PMID: 364941.

[4] K. A. Dill and J. L. MacCallum. The protein-folding problem, 50 years on. *Science (New York, N.Y.)*, 338(6110):1042–1046, Nov. 2012.

[5] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The KEGG databases at GenomeNet. *Nucleic Acids Research*, 30(1):42–46, 2002.

[6] E. Lolis and G. A. Petsko. Crystallographic analysis of the complex between triosephosphate isomerase and 2-phosphoglycolate at 2.5-a resolution: implications for catalysis. *Biochemistry*, 29(28):6619–6625, July 1990. PMID: 2204418.

[7] R. Machiraju, W. Ray, and C. Bartlett. BioVis data contest, 2013.

[8] M. Prabhakaran. The distribution of physical, chemical and conformational properties in signal and nascent peptides. *Biochem. J.*, 269(3):691–696, 1990.

[9] K. Tomii and M. Kanehisa. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Engineering*, 9(1):27–36, Jan. 1996. PMID: 9053899.

[2] http://www.mu-8.com