# Bio+MedVis Challenge IEEE VIS Mouse-Human Hybrids

Hauke Bartsch* Laura Garrison† Stefan Bruckner‡

University of Bergen, Norway

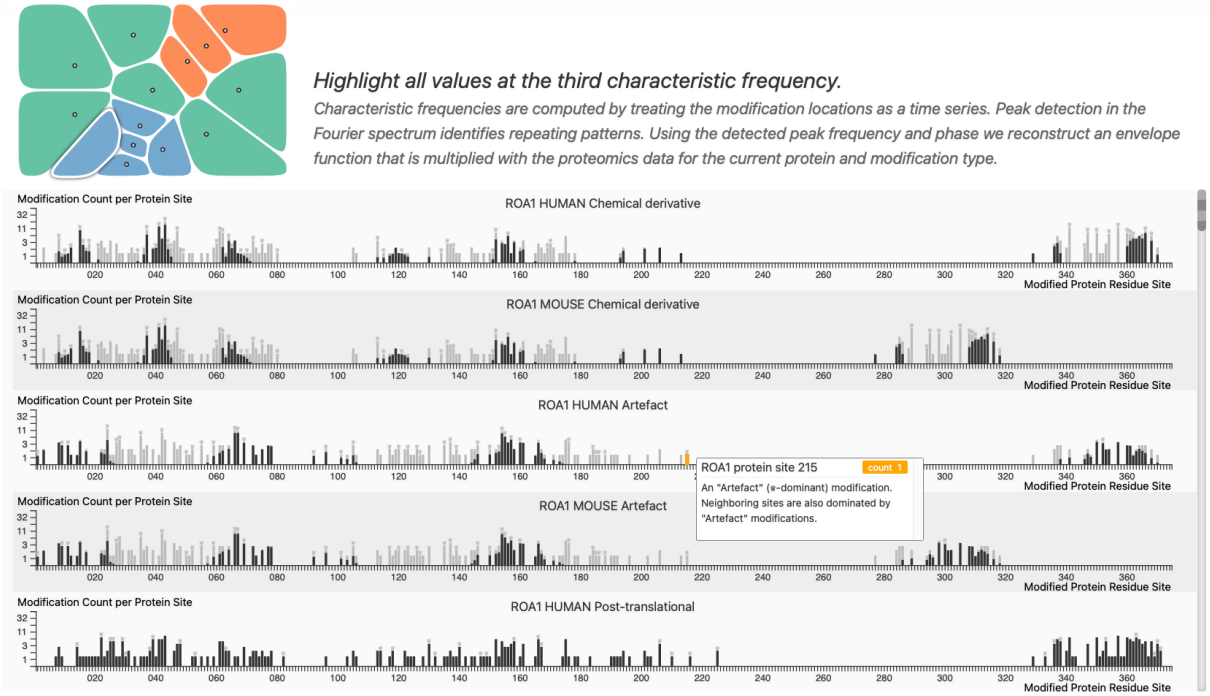


Figure 1: Voronoi treemap as a control surface and bar-plot animation facilitate identification and comparison of residues with high occurrences of chemical modifications for human and mouse models.

## Abstract

A key goal in the analysis of proteomics data is to compare chemical modifications occurring on different residues in a protein sequence between human and animal models to assess the suitability of animal models for drug development. Differences in modification patterns between mouse and human protein variants may also indicate evolutionary pressures to preserve functional regions distributed along the protein sequence. However, comparing and contrasting the many possible patterns of modifications is a significant challenge in visualizing proteomics data. For this year's Bio+MedVis Protein Beasts Challenge, we propose an exploratory interface for pattern discovery. Results from data-driven pattern detectors are displayed using an animation metaphor that is suitable for both comparing and contrasting tasks. In addition, we propose a novel, compact steel-drum like pattern selector component that abstracts from the details of the analysis method and focuses instead on the more playful aspects of an exploratory data analysis that favors instant gratification.

**Index Terms:** Human-centered computing—Visualization—Visualization techniques—; Human-centered computing—Visualization—Visualization design and evaluation methods Human-centered computing—Visualization—Visualization techniques—Treemaps

*e-mail: hauke.bartsch@uib.no
†e-mail: laura.garrison@uib.no
‡e-mail: stefan.bruckner@uib.no

## 1 Motivation

Proteins are the machines performing the basic functions of our cells and are an important target for medications. Evolutionary pressures in the mammalian line have preserved many aspects of proteins between animal models and humans. Together with cell repair mechanisms that prevent many accidental changes, any modifications we do observe are of interest, as they can equally reflect different functional variants or regions of the protein that are less important for its function. In drug development, this information helps in identifying drug targets can be used for predicting the success of translation studies from mouse to human. For this year's Bio+MedVis Challenge at IEEE VIS we propose an accessible exploratory visualization interface for pattern discovery. The application focuses on comparing and contrasting the residue modifications between human and mouse variants of protein sequences. The main goal of the visualization is to consistently present a larger number of detected patterns for exploratory data analysis supporting the user in understanding the data better by showing repeated patterns and differences between humans and animal models. The application abstracts from algorithmic complexity and the need to set up individual analysis workflows in favor of an accessible exploration of large scale patterns that inform future hypothesis-driven research. An interactive prototype is

available at https://tinyurl.com/yuxnabpm.

## 2 Proposed application

Our key requirements for data exploration include: **[R1]** Avoid sequential search by providing pre-attentive cues for pattern changes, **[R2]** Provide general methods for comparing and contrasting individual modifications between mouse and human variants, and **[R3]** Focus on a playful exploration of data-driven pattern visualization to encourage discovery.

The interface is separated into a control area for pattern selection at the top and a pattern visualization area below designed to be used on a desktop device.

Our application facilitates comparing and contrasting a pattern of modifications at each site of a given protein based on the familiar bar-style data visualization that displays locations along the protein on the x-axis and the number of modifications as height (see Fig. 1). This visualization does not take into account the spatial folding of the protein important for its function in favor of a complementary display of the sequence of protein modifications. We define a pattern as a salient sequence of scalar values that may highlight individual locations along the protein sequence. We map each pattern to motion using a cyclic animation of the height of each bar. Highlighted bars are moved more than bars that are not part of the selected pattern. Using motion to relay information allows for a quick perception of a large-scale pattern without the need for attention and multiple eye-movements in a visual search ([R1], [3], [2], [4]), as the human visual system is highly sensitive to motion. Our patterns of protein modifications become the tiger in the woods.

As part of this challenge, three proteins have been selected that each have a human and a mouse variant with multiple modification types. We expect different patterns to be present in each protein pair, so we select pattern detectors that are expected to be general enough and rely on data-driven analysis. We focus on (i) a set-theory based assessment of what mouse and human variants share and what makes them different, (ii) a quantile analysis of the modification magnitudes to highlight extreme values and (iii) a frequency analysis to identify repeated patterns in the sequence of modification locations [R2]. The set-based patterns visualize the union, intersection, and difference of modifications in each species pair, as set methods carry over modification information between species. In effect, we create mouse-human hybrid visualizations.

Whereas the pattern detectors in (i) are directly related to contrasting mouse and human variants of the same protein and modification type, the pattern detectors for (ii) and (iii) are specific to the individual proteins of human and mouse. Any pattern detected by them becomes apparent, because the interface groups human and mouse variants as pairs for each modification type. A scroll-bar allows the user to move between pairs of protein and modification type.

As a control interface for pattern selection we use a hierarchical area visualization similar to Voronoi tree maps of Balzer et al. [1] used in the visualization of software metrics. Such visualizations are effective in providing a compact, visually appealing, high density display where spatial arrangement and size can be used to support a varying number of hierarchically organized areas (touch zones) [R3]. The visualization is created by arranging the set-based pattern types (i) in a level-0 circle and the quantile (ii) and frequency analysis pattern types (iii) are added as further level-1 circles around their level-0 parent pattern. A weighted Voronoi tesselation with a square boundary area generates cells that visualize touch zones for each pattern detector type. Color is used to indicate level-0 (blue) and level-1 (orange) patters. To facilitate the playful data exploration, we adjust the tree maps and smooth the individual zones sufficiently to resemble a *steel drum* interface.

Whereas the control interface invites users to freely explore the patterns, more detailed textural information on each pattern generator is displayed next to the control area. When the user selects a pattern detector the text is updated, the pattern animation activates for all proteins and mutation pairs, and the user is free to inspect the pattern or scroll though the protein and mutation pairs.

We also added a second lighter-colored copy of the data to each bar-style plot indicating the start position of the bar animation. The darker bars show the end-point pattern of the cyclic animation at approximately .5 Hz. In Fig. 1 this indicates a level-1 pattern of characteristic frequencies detected in the ROA1 proteins for the modification of chemical derivatives and artefacts.

## 3 Conclusion

We propose a data exploration interface that allows for a quick inspection of spatial patterns and contrasts between mouse and human version of proteins. Our design integrates a larger number of data-driven pattern detectors by automatically computing candidate patterns that are displayed in a consistent manner.

Many algorithms are described in the literature for the detection of patterns and to contrast values in modification data for proteins. Our choice is not exhaustive or claims to cover the important aspects of this data resource. Instead we focused on sufficiently many pattern detectors to illustrate the suitability of motion to relay complex patterns in proteomics data and to show how a simplified interface can be used to integrate a variety of such methods. It remains to be shown if frequency based analysis methods of modifications correlate with data obtained from protein folding or the arrangement of active zones.

The Voronoi tree map visualization can utilize different weights for each zone to highlight the relative importance of pattern detectors at one level. A zone with a higher weight will have a larger relative area compared to a zone with a lower weight. In the current implementation this feature remains unexplored and initial tests have resulted in unsatisfactory results if this feature is used. This might be caused by the use of a circular arrangement of zones that inherently introduces size differences between center and surrounding zones.

Further improvements of the application are required to allow for a selection of protein and modification types to allow for an easier comparison of pattern across modification types. Detailed information on the individual modification per protein site could also be added in the form of tool tips.

This data exploration application is available in source code format at https://github.com/mmiv-center/human-mouse-hybrids.


## Acknowledgments

This research is supported by the University of Bergen and the Trond Mohn Foundation in Bergen (#813558, Visualizing Data Science for Large Scale Hypothesis Management in Imaging Biomarker Discovery (VIDI)). Parts of this work have been carried out in the context of the Mohn Medical Imaging and Visualization Centre (MMIV) and the Center for Data Science (CEDAS) in Bergen, Norway.


## References


[1] M. Balzer, O. Deussen, and C. Lewerentz. Voronoi treemaps for the visualization of software metrics. In *Proceedings of the 2005 ACM Symposium on Software Visualization*, SoftVis '05, p. 165–172. Association for Computing Machinery, New York, NY, USA, 2005. doi: 10.1145/1056018.1056041

[2] D. Fisher. Animation for visualization: Opportunities and drawbacks. *Beautiful visualization*, 19:329–352, 2010.

[3] A. P. Hillstrom and S. Yantis. Visual motion and attentional capture. *Perception Psychophysics*, 55(4):399–411, 1994. doi: 10.3758/BF03205298

[4] M. Waldner, M. Le Muzic, M. Bernhard, W. Purgathofer, and I. Viola. Attractive flicker — guiding attention in dynamic narrative visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2456–2465, 2014. doi: 10.1109/TVCG.2014.2346352

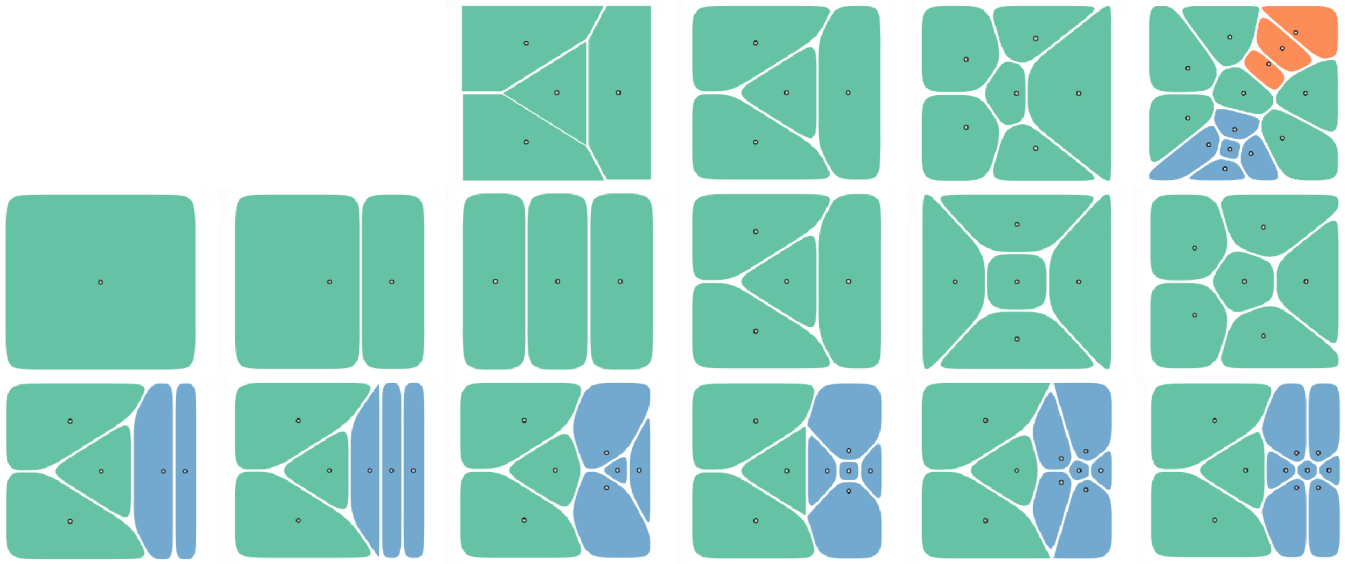

Figure 2: Control interface examples. The top row shows on the left a naive Voronoi tesselation (blue zones) of a square space with three control points placed around a central point (black circles). Following images show the same control points with a smoothed tesselation, a smoothed tesselation with a larger weight of the right-sided control point and, on the far right, the more complex arrangement of 14 control zones used in this work for pattern exploration. The middle row shows six level-0 tesselations (zone number one to six). In the bottom row one level-0 zone is sub-divided further into two to seven zones.

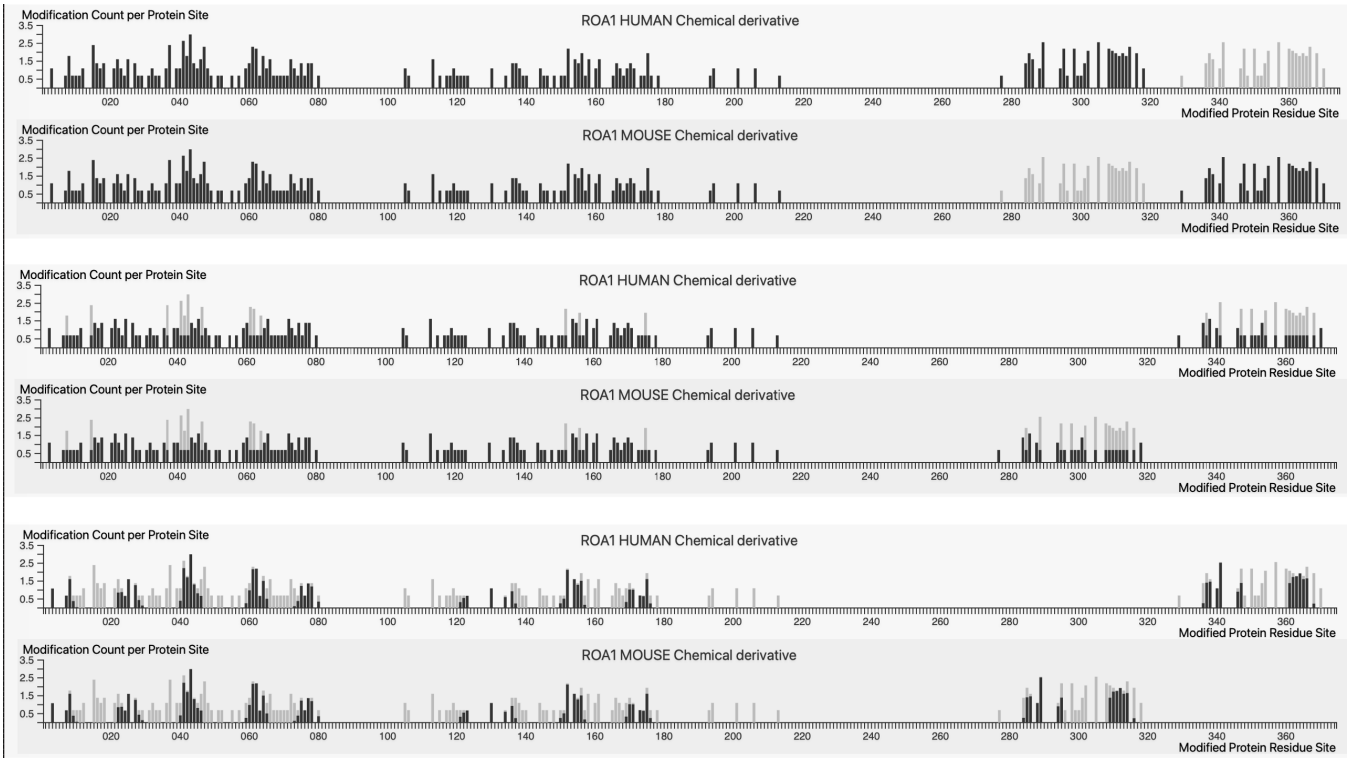


Figure 3: Example (static) pattern visualized with our application for pairs of mouse/human chemical derivative modifications in the ROA1 protein sequence. The top pair of plots show the baseline human (top) and respectively mouse (bottom) modification data in lighter gray. The paired mouse/human data are displayed in darker gray. Note that in the application the darker bars are animated, highlighting the additional contribution of the other species in the hybrid display. The middle pair of bar plots show the extreme value quantile pattern. The bottom row highlights one of the frequency analysis patterns.