DOI: 10.1111/j.1467-8659.2011.01949.x
Eurographics / IEEE Symposium on Visualization 2011 (EuroVis 2011)
H. Hauser, H. Pfister, and J. J. van Wijk
(Guest Editors)
Volume 30 (2011), Number 3

WaveMap: Interactively Discovering Features From Protein Flexibility Matrices Using Wavelet-based Visual Analytics

Scott Barlowe[1], Yujie Liu[1], Jing Yang[1], Dennis R. Livesay[2], Donald J. Jacobs[3], James Mottonen[3], and Deeptak Verma[2]

[1]Dept. of Computer Science, University of North Carolina at Charlotte, USA
[2]Dept. of Bioinformatics and Genomics, University of North Carolina at Charlotte, USA
[3]Dept. of Physics and Optical Science, University of North Carolina at Charlotte, USA

Abstract

The knowledge gained from biology datasets can streamline and speed-up pharmaceutical development. However, computational models generate so much information regarding protein behavior that large-scale analysis by traditional methods is almost impossible. The volume of data produced makes the transition from data to knowledge difficult and hinders biomedical advances. In this work, we present a novel visual analytics approach named WaveMap for exploring data generated by a protein flexibility model. WaveMap integrates wavelet analysis, visualizations, and interactions to facilitate the browsing, feature identification, and comparison of protein attributes represented by two-dimensional plots. We have implemented a fully working prototype of WaveMap and illustrate its usefulness through expert evaluation and an example scenario.

Categories and Subject Descriptors (according to ACM CCS): I.5.5 [Pattern Recognition]: Implementation—Interactive Systems

1. Introduction

Understanding the physical characteristics of protein structures is of great importance in developing new pharmaceuticals. Knowing how a protein will react under given circumstances will allow scientists to more finely tune drugs to specific diseases and genetic therapies to individuals. As many other fields in the biological sciences, researchers in this area increasingly rely on computational models to describe and predict physical phenomena. For example, protein prediction models have been used to understand the physical flexibility of local or regional subunits (residues) which alter overall protein structure. Since structure is the most important factor influencing protein behavior, insights from these models are crucial in predicting a protein's function.

Computational models can be quickly changed based on new domain knowledge or used to examine the effect of varying parameters. The large number of possible outputs often obscures important differences in model settings. It is these subtle and unexpected variations which can be the most important in relating model construction to a desired physical behavior. The number of different model outcomes is only limited by the number of variables encoded as model parameters. A single model output often reflects the behavior of several hundred residues per parameter set where each residue may be responsible for influencing overall structure and function. Clearly defining the relationships among residues for one model output in the midst of perhaps hundreds of different parameter set combinations is a challenging problem. Methods that aid in clearly defining the relationship between residue flexibility parameters and model outputs will greatly advance biomedical knowledge.

In this paper, we present a novel approach in applying visual analytics in an effort to aid scientists in the browsing and understanding of large-scale two-dimensional flexibility data resulting from a protein prediction model. Our work integrates automatic analysis and coordinated visualizations to help guide domain analysts to important features hidden in the data. In the following sections, we first provide background knowledge in protein structure and function. Second, we describe a currently used protein prediction model and list the necessary high-level visualization tasks for effective model usage. Third, we present related work from the image analysis and the visualization communities. Fourth, we present our visual analytics approach that connects feature extraction techniques with a highly interactive interface to address the identified high-level tasks. Finally, we show the

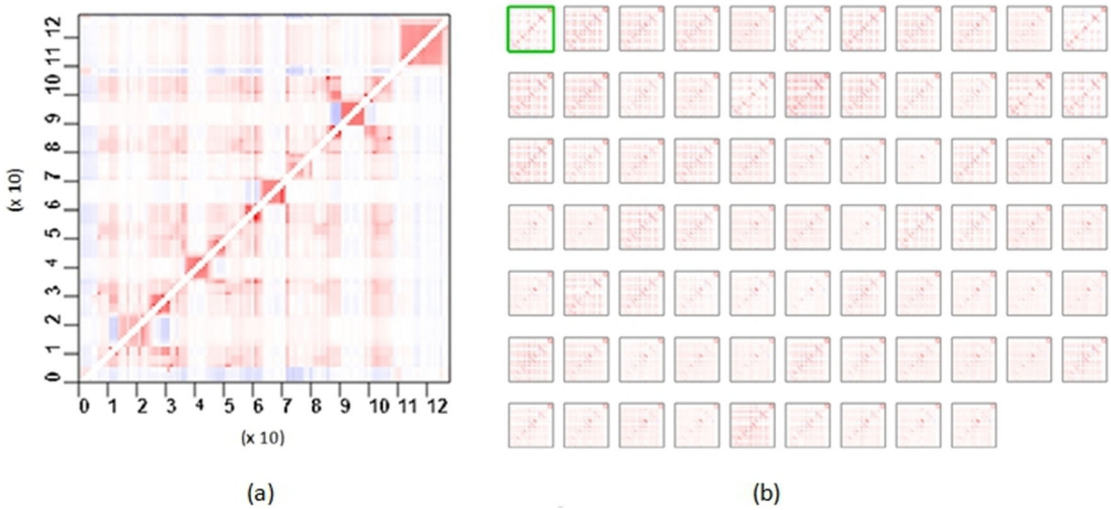

Figure 1: *(a) A typical flexibility-response plot showing allosteric response for the CheY protein [MJL10]. A color index at i, j is the response of residue j occurring due to a perturbation at residue i. (b) Analyzing subtle patterns within flexibility measures for a single plot are difficult and placing them in context makes the problem worse. The plot to the left is highlighted in green.*

effectiveness of our approach through expert evaluation and an example scenario.

2. Protein Structure and Function

Proteins are composed of linear chains of amino acid building blocks connected by chemical bonds [PSW*06]. These polymers can consist of 20 different possible amino acids (residues) and vary greatly in length and sequence across different proteins. The spatial arrangement of these residues determines the biological function of the protein. Although understanding the relationship between molecular structure and function is important, there are few global scale connections that correlate specific structural changes to function [KJB00].

One of the main obstacles in understanding changes in proteins is the vast number of possible spatial configurations that may occur under numerous environmental conditions. This set of possible configurations is referred to as a protein's conformation space [NHL07]. Most prediction methods search this space for the structure having the lowest energy (an indicator of structural stability), making energy one of the primary variables to be examined. Although important in protein changes, the energy function is often difficult to explore. Exploration is complicated by multiple local minima and an energy function's dependence on many interrelated terms. Furthermore, differences in the underlying chemical structures complicate analysis.

3. The Distance Constraint Model and Allostery

Because of the complexities associated with exploring a protein's conformation space, researchers rely on models to understand and predict changes in protein structure. The Distance Constraint Model [JD05], [LDWJ04] is an example of a model that has proved helpful in understanding protein function. The Distance Constraint Model (DCM) is based on free energy decomposition and mechanical constraints. The premise of the DCM is to relate free energy and mechanical constraints with a graph topology which can be calculated in linear time. Like any model that retains relevance, the DCM is constantly evolving and the need to quickly examine changes in parameters is crucial for efficient model development. The DCM outputs a large set of mechanical and thermodynamic descriptions of the target protein.

The data resulting from the Distance Constraint Model [JD05], [LDWJ04] are flexibility indices which quantify the ability of a residue to change its structure and, as a result, its function. One example use of flexibility indices is the measurement of allosteric response. Allosteric response measures how a change (perturbation) in one residue may induce flexibility changes in other residues [MJL10]. In many cases, the perturbed residue may not be spatially adjacent to the responding residue. Furthermore, the resulting change may require a coordinated effort among multiple residues spanning the entire protein. Being able to investigate and compare the different allosteric responses for many parameter sets will allow scientists to more speedily recognize biochemical mechanisms necessary to achieve a desired biomedical result.

Flexibility changes for allosteric response can be pre-

sented as an *n x n* asymmetric color plot where *n* is the number of residues that make up a protein. The residues that comprise the protein are ordered according to the three-dimensional residue sequence giving significance to both local and regional characteristics. A typical flexibility-response plot [MJL10] is shown in Figure 1(a) for the 128-residue CheY protein, revealing the protein's change in flexibility for a given set of model parameters for the DCM [JD05], [LDWJ04]. In this plot, a color index at *i, j* is the expected flexibility change of residue *j* occurring with a perturbation at residue *i*. Values are normalized between -1 and +1. Values less than zero result in increasingly darker shades of blue and indicate increased rigidity. Values greater than zero result in increasingly darker shades of red and indicate increased flexibility. White areas indicate no change.

Figure 1(b) shows the flexibility-response plots of the CheY protein for different sets of model parameters. Examining just a few plots is difficult and comparing hundreds of them proves to be a daunting challenge. In this case, there are over 6 million data points (128 residues x 128 residues x 75 parameter sets) of often small, subtle changes that need to be analyzed. Patterns that emerge within and among flexibility plots give insight into the mechanisms important for biological function. Interesting patterns include the size and location of similar or dissimilar regions, and any outliers where a certain residue may unexpectedly differ from its neighbors or from a consensus over the parameter space. The various types of DCM outputs, when taken together, provide both local and global descriptions of protein dynamics.

4. Tasks

The work presented in this paper has been motivated by the data analysis needs of the BioMolecular Physics Group (BMPG) in the Department of Bioinformatics and Genomics at UNC-Charlotte. The researchers in this group have been conducting extensive research in protein prediction models, primarily the design, development, and evaluation of the DCM [JD05], [LDWJ04]. Although the DCM has proven to be a powerful computational model, only a very small portion of the DCM can be used at once because of the limited capability of analyzing the flexibility plots. Usually, analysis consists of manually examining one plot at a time. The researchers expect that a visual analytics tool that allows them to effectively conduct the tasks on a much larger scale will greatly advance their understanding of how the DCM predicts protein flexibility.

To identify the visualization needs, the visualization experts have observed the weekly BMPG meetings on a regular basis for over two years. During this time span, the visualization experts and the BMPG researchers also conducted more than 25 face to face meetings to discuss the visualization needs of DCM scientists and to provide feedback for the visualizations developed. According to the long-term observation and continued collaboration, several high-level visualization tasks of DCM researchers when using flexibility plots have been identified. The tasks are presented below to explain the motivation for the design of our visual analytics tool.

- **Analyzing spatial relationships and numeric trends of flexibility measures within proteins.** Protein dynamics can be altered by either a local group of residues, larger regional groups, or the concerted effort of multiple areas of varying sizes. Locating and identifying those regions of interest which contribute to change is necessary before the roles of individual subunits can be identified. For example, a region with an isolated change in flexibility is highlighted in the example scenario (Figure 2). This area can be studied in greater detail so that it can be compared to other regions within the same protein. Any differences within a single protein can explain why certain areas have a greater propensity for change and result in a given overall structure.
- **Studying parameter influence and grouping parameter sets.** Parameter refinement within a model is a reflection of evolving expert knowledge for a specific protein and environmental condition. For a fixed parameter set, a comparative analysis between different proteins and/or environmental conditions can help discover new spatial relationships and numerical trends. Grouping model outputs by parameter sets will allow scientists to understand what combination of parameter settings result in the greatest or most unexpected change for a single or group of residues. From this knowledge, domain experts can refine the model or investigate ways to take advantage of these differences.
- **Pruning parameter sets and residues** Clearly defining relationships among residues and parameters of interest is best accomplished if redundant or uninteresting data items are excluded from consideration. This can take the form of excluding entire parameter sets, entire proteins, or individual residues based on domain knowledge or thresholding. Additionally, scientists need to be able to start with a well-studied individual residue, group of residues, or overall structure that is accurately reflected by model outputs and then eliminate parameter sets based on similarity (or dissimilarity) from the established item.

According to BMPG members, the above tasks are significant and frequently encountered. For example, *clamping* is an application where these tasks are needed. Clamping is the restriction of a single residue's flexibility. As the clamp is moved systematically along the protein's structure, flexibility is recalculated. The scientists must compare the flexibility trends for specific regions before and after each clamp. Researchers would then like to see how similar the responses are at different clamp positions, and for different model parameters. Unexpected clamp responses among multiple plots will be isolated for further analysis.

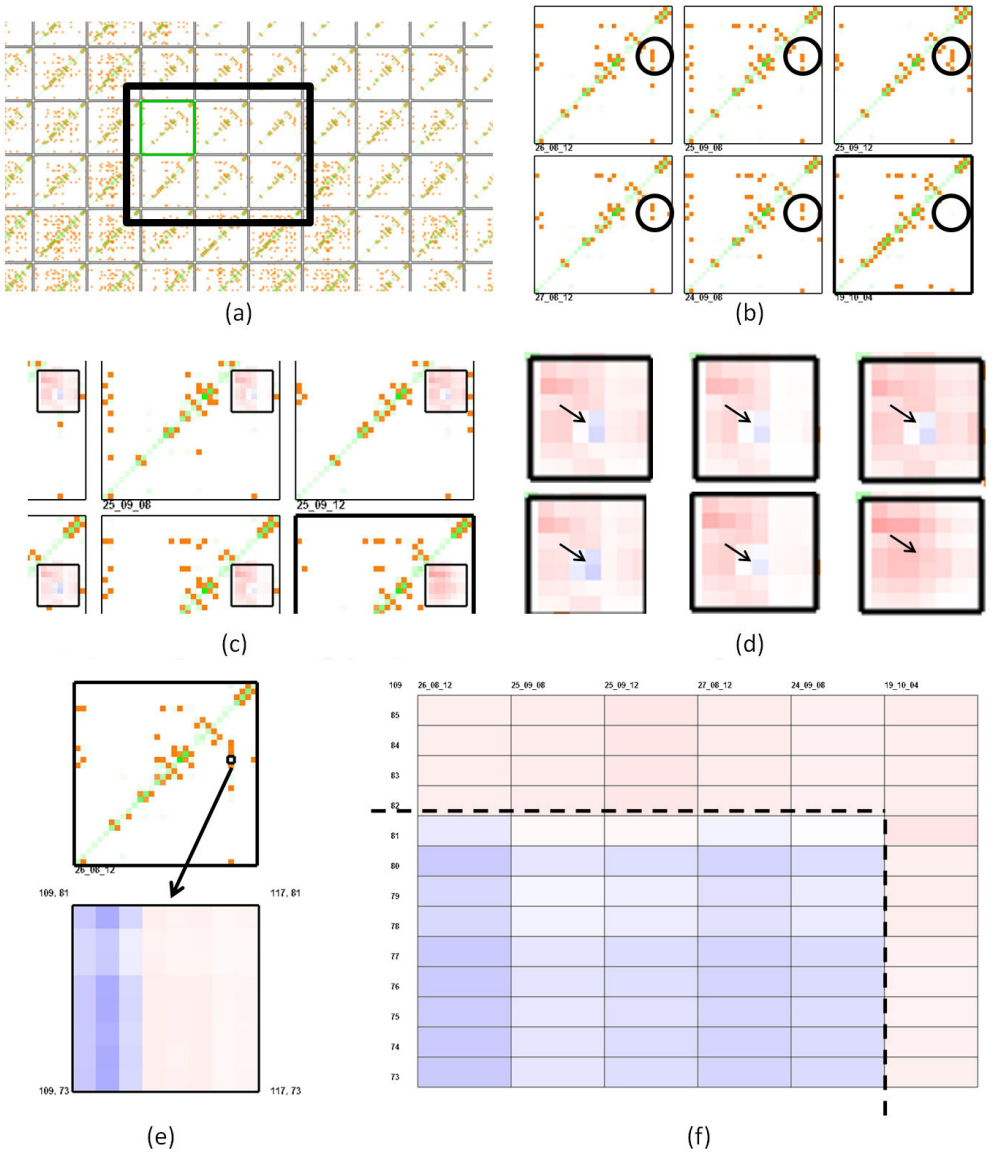

Figure 2: *(a) A section of the Jigsaw layout is identified for further analysis based on similar global features (2 decompositions). A known plot is highlighted in green. (b) Closer examination of features reveals noticeable differences. Circled regions indicate a point of difference in one parameter set (lower right). (c) The coordinated lens reveals the pattern that the features emphasize. (d) A larger view of the lens shown for clarity. (e) Plot coordinates for the region of interest are found during reconstruction. (f) Detailed analysis allows further mapping of flexibility values to residue numbers among multiple plots.*

5. Related Work

5.1. Visualization

Currently, scientists lack effective tools to conduct the above tasks for large flexibility data sets. Existing methods heavily depend on manual inspection of enlarged flexibility plots using Heatmaps [MXJL09], [MJL10]. Subtle but important relationships and patterns may remain hidden even with zooming and distortion interactions. The large number of plots and the subtle differences both within each plot and among parameter sets are almost impossible to distinguish. These obstacles greatly hinder knowledge discovery. To the best of our knowledge, our work is among the first efforts from the visualization community addressing entire flexibility data sets and automatically guiding users to potential areas of interest. Note that we chose to keep the current color

and arrangement scheme for visualizing the original flexibility values since they are ingrained into current workflows and convey important spatial information.

Heatmaps are widely used in bioinformatics besides protein flexibility data visualization. Specifically, they are the most common representation for gene expression data [GOB*10]. Many of them can guide users to patterns such as clusters and outliers within the data. For example, HCE [SS02] and Java Treeview [Sal04] enable users to identify clusters in microarray experiment data sets using hierarchical clustering algorithms and interactive visual exploration. Visualization methods developed for matrix data [TLKT09], [SM07] also allow users to find patterns through interactive visual exploration. However, the above techniques would not work for protein flexibility data visualization because of their heavy reliance on similarity grouping. Not only does each *i,j* value have a color-coded flexibility measure but also carries spatial significance reflecting residue ordering along the three dimensional protein structure. Any reordering or rearranging of rows, columns, or individual measures would disrupt the spatial context in which any flexibility measure occurs. Moreover, a large number of plots need to be examined simultaneously in our application while most of the above techniques only consider one matrix or array at a time.

5.2. Image Analysis

Many image analysis techniques for extracting features can be applied to flexibility plots so that difficult to detect patterns can be identified. For example, Principal Component Analysis (PCA) [Jol02], [Dun89] can be used to summarize features by finding the linear combinations of variables and then ordering the resulting components by variance [YR01]. It has been used in many image processing applications such as face recognition [PKP10] and edge detection [QPA08] but can be computationally taxing [VRP09], [EYND10]. Additionally, the results can be difficult to interpret [ZHT06]. Fourier analysis [GW01] is another popular image analysis technique applied to many areas including feature extraction and dimension reduction [JH07]. Fourier frequencies can be linked to pixel value changes where low frequencies are associated with slowly varying pixel changes and high frequencies are associated with abrupt pixel changes [GW01]. A major drawback to this type of analysis is that frequency and spatial information cannot be conveyed at the same time.

Our approach uses wavelet lifting [Swe96] for automatic feature extraction. Wavelet analysis is based on small signals (waves) of limited duration and varying frequency [GW01]. This type of transform allows the same frequency-based processing of pixel values as Fourier analysis. However, wavelets provide simultaneous frequency and spatial information with a multi-resolution approach that allows normally hidden features to be revealed. Wavelet lifting [Swe96] is an improvement to traditional wavelet analysis. An iterative process which uses no more memory than required for the original data matrix improves computational efficiency. Additionally, the results of lifting can be reversed by simply reversing the discrete steps used in transformation [JCH01]. This property enables the original data to be directly and quickly accessed from any level of decomposition. Wavelets have been integrated into visualization tools for brushing applications [WB96], text analysis [MWBF98], and many scientific applications [FHH03], [Bon97].

We chose wavelet lifting because of its capability to provide simultaneous frequency and spatial information in a multi-resolution approach, the ease of reversing the transform, and efficient implementation. Other approaches are possible and we plan to explore them in the future.

6. WaveMap

We propose WaveMap (Figure 2), a visual analytics approach that integrates wavelet lifting [Swe96] with visualization to address the needs mentioned in section 4. It consists of an overview to display the entire data set, a feature window to examine selected plots, and a detailed analysis window to perform closer examination. In the overview (Figure 2(a)), features extracted by wavelet lifting are visually presented to users to help them locate global trends or local areas where trends are interrupted. Subtle trends that are hard to discover in the original data become visible in the feature space. To study parameter influence, a set of plots in the original data or extracted features can be viewed in a clustering or sorting layout from which the global trends across parameter sets, as well as clusters and outliers of plots can be observed. Users can interactively retrieve groups of interesting plots so that further examination and comparison can reveal the relationships among residues and parameters (Figure 2(b)). Features can be filtered based on their type and magnitude and are intuitively mapped to the original data (Figures 2(c) and (d)). The feature window allows examination of interesting regions for a subset of interesting plots (Figure 2(e)). The detail window facilitates coordinated, residue-level analysis among multiple plots (Figure 2(f)). In every view, specific regions for given parameter sets can be exported for insight management and exchange.

6.1. Feature Extraction

Wavelet lifting is first applied to extract varying plot features at many resolutions. During each application to a discrete, two dimensional signal (or ***decomposition***), the data is separated into the high and low frequency components. The result is a series of four data matrices each of which is one-quarter the size of the original data matrix. The results include the high frequency components in both directions (HH), low frequency components in both directions (LL), low frequency along rows and high frequency along columns (LH), and high frequency along the rows and low frequency

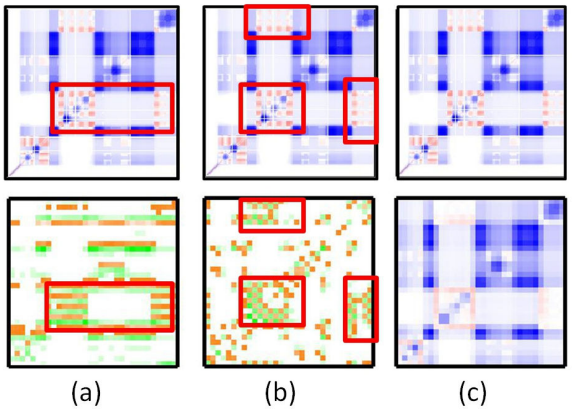

Figure 3: *Three cooperativity correlation plots [LJ06], [MXJL09] depicting the correlation of flexibility changes. Original correlation values are shown on the top row. Blue indicates co-rigid regions and red indicates co-flexible regions. The data after wavelet transformation (3 decompositions) are on the bottom row and corresponding regions are bounded in red. Orange highlights positive correlation changes and green highlights negative correlation changes in all except the last pair. (a) Row patterns are preserved while indicating changes along columns. Because this data set is symmetric, the LH and HL subbands are redundant. Only the LH subband is shown here. (b) Areas of change along both rows and columns are detected. (c) Coarse-grain characteristics are preserved for the entire data set.*

along columns (HL). One of the components is chosen to be fed to the input of the next stage and the process repeats.

Wavelets come in varying families and can be designed to extract desired features. We use a lifting implementation of the widely-applied Debauchies 4 wavelet [JCH01] but other wavelets can also be used. Because the data is halved along the rows and columns during each application, we linearly interpolate the original data set so that each row and column is a power of two. The interpolated data is only used in application of the wavelet algorithm and is never visible to the user. After each decomposition, any given feature represents a larger neighborhood in the original data. Feature magnitude can be mapped to color so that the degree of change between adjacent locations can be visually represented. The components, or ***subbands***, and the characteristics emphasized in the data that are important for our work include

- **LL:** Averages along rows and columns preserve regional residue responses.
- **LH:** Averages along rows and change detection between adjacent columns preserve residue behavior along rows while revealing differences among neighboring residues along columns.
- **HL:** Averages along columns and change detection between rows preserve residue behavior along columns while revealing differences among neighboring residues along rows.
- **HH:** Change detection along rows and columns reveal differences in residue response along both rows and columns.

It may seem that the total amount of data has been significantly increased because each original data plot is now represented by four different subbands. However, each decomposition results in each subband being only one-quarter of the input data size. Additionally, the subbands and multiple levels of resolution produced are different perspectives of the original data. This allows experts to choose the appropriate prism through which domain knowledge can be applied.

Figure 3 illustrates how features for each subband emphasize various characteristics present in cooperativity correlation flexibility plots for a TRX protein [MXJL09]. The data is slightly different from the allosteric response in that each index is a correlation measure between two flexibility measures. The top row is the original data and the bottom row is the transformed data after wavelet analysis and filtering. Original plot values (top row) are displayed with the red-white-blue scheme currently in use by domain scientists. The features for all subbands except for the averaging features (LL), use a different color scheme because features capture the *change* occurring between plot regions and not the original plot values. We chose a feature color scheme ranging from green (negative changes) to white (no change or below a threshold) to orange (positive changes). The red-white-blue scheme was kept for the averaging features because they visually relate original data information at varying resolutions (Figure 3(c)). In Figure 3(a) coarse-grain behavior along rows is preserved and differences along each column are detected so that general residue behavior is preserved along the rows but changes in residue behavior between adjacent column are detected. Figure 3(b) shows the detection of changes along both rows and columns revealing where residue behavior changes from adjacent residues in both the row and column direction. Figure 3(c) preserves course-grain behavior in both directions. Although the features highlighted in the top row are easily detected and represent symmetric data, such features can be much harder to detect in other datasets without the help of the transformed data. We will show in Figure 2 how the transformation applied to the asymmetric and much more subtle allosteric response data can ease analysis.

We briefly introduce a formal user study that confirmed the effectiveness and efficiency of the feature displays in guiding users to identify and compare subtle features in flexibility plots. Eight student subjects (no domain knowledge was required in the tasks) conducted the study one by one. The task was a matching game that required the subjects to find a flexibility plot in a set of 12 plots that had the most

Figure 4: *Clustering features for a section of the (a) MDS and (b) Jigsaw layouts. The HH subband after 2 decompositions identifies places of change along both rows and columns while simultaneously reducing the number of data points. In both configurations, the data is separated into areas with plots having many points of change (left side of each layout) and plots with fewer points of change (right side of each layout).*

similar patterns to a given plot in the highlighted sub-region. The subjects conducted the task with and without the feature displays and the order was balanced. The results showed that all the subjects conducted the task successfully with the help of the feature displays while only one subject succeeded without the feature displays, even when enlarged flexibility plots were provided. The subject rating after the study revealed that the perceived difficulty was much greater without the feature displays and subject confidence that they had chosen the correct plot was higher with the feature displays.

6.2. Overview

Wavemap allows users to explore a large number of original or transformed plots in an overview. The positions of the plots in the overview visually reveal their similarities. Users can interactively set similarity measures, which can be the covariance between the original plots or desired subbands at the level of decomposition of interest. When clustering is based on the covariance of features, plots sharing common subband features are considered similar.

The overview provides three layouts to reveal the similarities among the plots:

- **Multi-Dimensional Scaling (MDS) layout:** MDS allows viewers to interpret overall plot similarity as visual distances [BG09], [CC00]. In the MDS layout, similar plots are close to each other while dissimilar plots are far away from each other (Figure 4(a)). When parameter set labels are turned on, analysts can see if a relationship exists between any parameter sets and if any parameter combinations result in outliers. Plot size can be interactively reduced to minimize overlap.
- **Jigsaw layout:** The jigsaw layout [Wat05], [YHW*07] allows users to examine plot clusters and outliers without overlap. It is a grid layout where similar plots are placed close to each other and boundaries among clusters of similar plots can be detected (Figure 4(b)).
- **Sorting layout:** The plots are ordered by their similarities to a selected plot so that users can examine their similarities in relation to the selected plot.

Users can display any subbands through a drop-down box and any decomposition level through a slider. They can also select similarity measures independent from how the plots are displayed. This flexibility allows the user to explore many clustering and display configurations based on desired characteristics. Users can manually select plots or automatically select plots whose similarities to a given plot are larger than a threshold the users set up. The selected plots are placed on a clipboard and carried to other views for detailed exploration.

6.3. Feature Exploration Window

Features can be examined in detail for plots placed on the clipboard in the feature exploration window. WaveMap provides a coordinated lens so that users can associate features with the corresponding original values without extra eye movement. When users sweep the lens over the features, the averages (LL) are displayed at the current decomposition level in the lens to associate features to the underlying original data (Figure 5(a)). Lenses are displayed on all plots under exploration at the same position. Whenever users move the lens over one of them, all the other lenses will also move to the corresponding position in their plots to allow comparison. The users can interactively change the size of the lens.

Reconstruction further extends the effective and efficient association of extracted features with the original values (Figure 5(b)). We allow users to select a plot from the clipboard as the focus of the feature exploration window and inspect its feature plot and original plot side by side. Users can interactively navigate in the feature plot by moving a bounding box using directional buttons or a snap function that automatically positions the box to the nearest feature with a magnitude above a threshold. Once a feature is accessed,

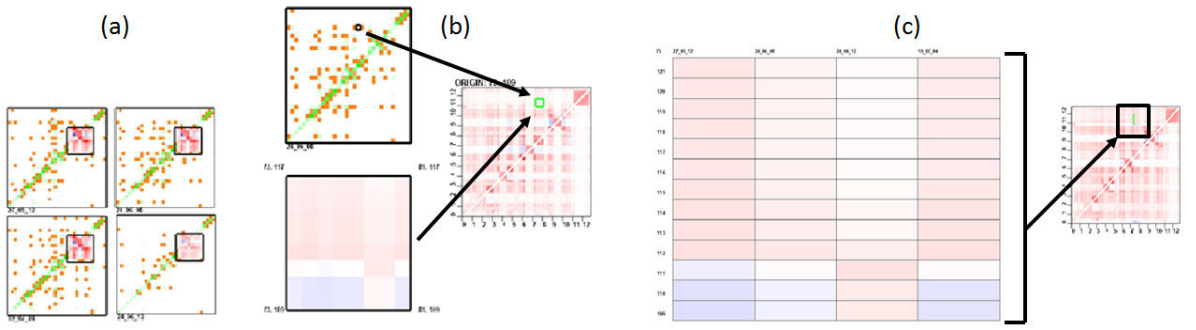

Figure 5: *(a) A coordinated lens allows simultaneous examination of multiple plots while relating features to the original data. (b) Reconstruction further bridges the feature and original data. Coordinates are marked in the large plot showing the original data and in the context window. A bounding box marks where the features occur in the data before transformation. Coordinates are propagated to the detail view. (c) Detailed analysis occurs for a column section across multiple plots. Left, right, up, and down buttons facilitate navigation during reconstruction and detailed analysis.*

the original data values responsible for that feature value are bounded in the original plot and in a resizable zoom-in display of the original plot. Plot coordinates of the bounding box are displayed at the corners of the zoom-in display. Context of the feature being examined is maintained by a smaller window showing the entire plot. Within the context window a green rectangle shows the bounds of the visited features.

6.4. Detail Analysis Window

Plots on the clipboard are also available for detailed analysis (Figure 5(c)). In this view, a column segment is shown for each plot. Each segment is composed of color rectangles that represent the changes in flexibility for residues in the column. The residue number for the current column is shown in the top left of the window and row numbers are shown to the left of the series of cells. Exploration in this view can begin at the plot origin or from the coordinates propagated from the feature exploration window. As during feature exploration, a window showing an entire plot of interest accompanies the detailed analysis window. A green bar indicates the current residue segment being displayed. Residue columns are navigated by directional buttons or by entering coordinates into text fields so that a detailed sweep across the plots can be performed. Clicking in a column makes the corresponding plot the focus of the context window. Removing an item from the clipboard removes it from the feature exploration and detail analysis windows and allocates the extra space to the remaining plots. Normalization to the range between -1 and +1 occurs only for plots in the display so that column segments can be more accurately compared. Users can also display residue name abbreviations within the segments to further connect residue sequence information with flexibility values. Users can save the residue segments under exploration into an image file for records and insight exchange.

7. Evaluation by Experts

The BMPG researchers have taken an active part in this work from the early requirement analysis stage to the design, development, and evaluation of Wavemap. Before any significant visual analytics approaches, views, and interactions were added into Wavemap, their potential usage was discussed with the BMPG researchers to make sure that they fit the high level needs of the domain experts. After a view or a major interaction became live in the system, feature inspections were conducted. In an inspection, the new view or interaction was demonstrated to the BMPG researchers and they examined its usability and discussed scenarios in which the features can be used in their applications. Features were often refined according to feedback from the inspections. The BMPG researchers are quite excited by the current features supported in Wavemap, especially the wavelet analysis guided feature inspection, the overview that allows them to evaluate their parameter set grouping hypothesis, and the multi-resolution approach that allows them to efficiently identify patterns of interest from a large data set and quickly dive into fine details.

8. Scenario

We now present an example scenario that illustrates how WaveMap can be utilized for better understanding of flexibility data. It specifically highlights the utility of our approach in detecting a small but significant area among a set of similar allostery response plots. A good reason for performing such analysis is to find model parameter sets which result in similar overall response but exhibit a small difference which could explain subtle variations in behavior. The data set used in the example represents flexibility response measures from the CheY protein.

A protein scientist pre-selects a set of parameters (ie a

plot) which exhibit a desired global behavior. The analyst suspects that other parameter sets globally similar to the selected plot have differences which may explain subtle variations in behavior. However, the values and residue region resulting in this behavior are unknown. The analyst searches for the known plot in the overview by entering the identifier into a search box and it is highlighted in green. Places of flexibility trend changes within this parameter set indicating possible differences are difficult to locate in context of other similar plots. The analyst moves the slider which changes the level of decomposition and examines the resulting features (Figure 2). When the wavelet features are displayed after two levels of decomposition and after the elimination of features having a small magnitude, the analyst sees a pattern of interest within the known plot indicating the changes in flexibility. He/she uses the Jigsaw layout [Wat05], [YHW*07] and identifies several adjacent plots with similar features to the known plot to ensure global similarity. The group of plots are placed on the clipboard for further examination.

After pruning the parameter sets, the analyst proceeds to the feature exploration window. In this window, the analyst easily compares corresponding regions of the selected plots with the help of the coordinated lens (Figures 2(b)-(d)). It is confirmed that the feature plots are similar but have noticeable differences. Of particular interest is the feature present in all but the bottom, far-right plot in Fgure 2(b). The coordinated lens (2(c) and 2(d)) reveals that the area of interest highlights a sharp change in flexibility (a small blue section in the middle of red) except for the one parameter set.

To more accurately define the region of difference, the location is visited (Figure 2(e)) and the residue numbers marking the area of change are revealed. The coordinates are propagated to the detail analysis window (Figure 2(f)) and the differences in response can be mapped to the specific residue numbers. The analyst can now further examine the physiochemical properties found in the plot lacking the small blue area to see if this region is perhaps responsible for variations in behavior or if the model should be modified.

9. Discussion

9.1. Extensions

The initial purpose of the proposed approach was to help computational biologists investigate outputs from the Distance Constraint Model [JD05], [LDWJ04]. As the approach was refined and the data set was better understood, we realized that biologists often need to visually describe data in context of spatial arrangement. Our approach helps locate, identify, and evaluate both local and regional factors influencing global behavior. However, any data set where the items can be arranged in a row/column layout that remains consistent across data items will benefit from our approach. If the data is in matrix form where order has spatial or other meaning, the interactions between variables can be revealed. The index *i, j* would represent the value of an output *i* for a parameter *j*. Each row (or column) is useful for evaluating many model or experimental responses for a single variable and each column (or row) reflects a single result for the combination of interdependent variables. Other domains that would benefit include microarray analysis [Liu09], economic forecasting, engineering simulations, and many more.

9.2. Limitations

During experimentation a couple of limitations were found. First, the effectiveness of this approach is influenced by the wavelet used. Each wavelet feature is the result of a correlation between two signals (data signal and wavelet signal) [LLZO02] and changing the wavelet alters its correlation to the data signal. This means that, for a different wavelet, the interpretation of the extracted features in relation to the underlying data may be different. Our initial implementation relied on a well-known wavelet technique and serves only as an initial attempt. Refinement and study are needed to develop custom wavelets so that more specific feature sets can be targeted and better defined in terms of biological significance. Second, our technique is meant for identifying localized neighborhoods which cumulatively result in a feature and not a precise data point. Finding a precise data point of interest may take additional exploration because of shifting during transformation and artifacts from noise. Boundary conditions can also lead to inaccurate representation of the underlying data near plot extremities and is an active area of research in signal processing [CA04], [GS00].

10. Conclusion

We have introduced a novel approach designed to ease the location and navigation of protein flexibility plots. WaveMap leverages strengths from signal processing and visualization in a single, coherent environment. Our initial evaluation with domain experts suggest that Wavemap is promising in supporting high level flexibility data exploration tasks such as analyzing spatial relationships and numeric trends, investigating parameter influence and the grouping of parameters, as well as pruning parameter sets and residues. Long-term user studies with domain experts are being planned to test our approach in depth. In addition, we would like to apply our approach to entire protein families and to other domains, both related and non-related to biological data. Experimentation with other feature extraction techniques, including custom-designed wavelets, is also of great interest.

Acknowledgements

This work was performed with partial support from NIH grant R01 GM073082 and the DHS Visual Analytics for Command, Control, and Interoperability (VACCINE) Center of Excellence, under the auspices of the Charlotte Visualization Center.

References

[BG09] BORG I., GROENEN P.: *Multidimensional Scaling: Theory and Applications.* Springer, 2009. 7

[Bon97] BONNEAU G.-P.: An introduction to wavelets for scientific visualization. *Scientific Visualization Conference (dagstuhl '97).* (1997), 16. 5

[CA04] CAI W., ADJOUADI M.: Minimization of boundary artifacts on scalable image compression using symmetric-extended wavelet transform. *International Conference on Information Technology: Coding and Computing 1* (2004), 598 – 602. 9

[CC00] COX T., COX M.: *Multidimensional Scaling.*, second ed. Chapman and Hall, 2000. 7

[Dun89] DUNTEMAN G. H.: *Principal Components Analysis.* Sage Publications, 1989. 5

[EYND10] ELSHENAWY L., YIN S., NAIK A., DING S.: Efficient recursive principal component analysis algorithms for process monitoring. *Ind. Eng. Chem. Res. 49*, 1 (2010), 252–259. 5

[FHH03] FARIN G., HAMANN B., HAGEN H.: *Hierarchical and Geometric Methods in Scientific Visualizations.* Springer, 2003. 5

[GOB*10] GEHLENBORG N., O'DONOGHUE S., BALIGA N., GOESMANN A., HIBBS M., KITANO H., KOHLBACHER O., NEUWEGER H., SCHNEIDER R., TENENBAUM D., GAVIIN. A.: Visualization of omics data for systems biology. *Nature Methods. 7* (2010), S56–S68. 5

[GS00] GUTIERREZ A., SOMOLINOS A.: Influence of wavelet boundary conditions on the classification of biological signals. *Proceedings of the IEEE 26th Annual Northeast Bioengineering Confernece* (2000), 25–26. 9

[GW01] GONZALEZ R., WOODS R.: *Digital Image Processing.* Prentice Hall, 2001. 5

[JCH01] JENSEN A., COUR-HARBO A. L.: *Ripples in Mathematics.* Springer, 2001. 5, 6

[JD05] JACOBS D., DALLAKYAN S.: Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophys J 88* (2005), 903–915. 2, 3, 9

[JH07] JADHAO D. V., HOLAMBE R. S.: Feature extraction and dimensionality reduction using radon and fourier transforms with application to face recognition. *International Conference on Computational Intelligence and Multimedia Applications. 2* (2007), 256–260. 5

[Jol02] JOLLIFFE I.: *Principal Component Analysis.* Springer, 2002. 5

[KJB00] KESKIN O., JERNIGAN R., BAHAR I.: Proteins with similar architecture exhibit similar large-scale dynamic behavoir. *Biophys J 78* (2000), 2093–2106. 2

[LDWJ04] LIVESAY D., DALLAKYAN S., WOOD G., JACOBS D.: A flexible approach for understanding protein stability. *FEBS Lett 576* (2004), 468–476. 2, 3, 9

[Liu09] LIU Y.: Wavelet feature extraction for high-dimensional microarray data. *Neurocomputing* (2009), 985–990. 9

[LJ06] LIVESAY D., JACOBS D.: Conserved quantitative stability/flexibility relationships (qsfr) in an orthologous rnase h pair. *PROTEINS: Structure, Function, and Bioinformatics 62* (2006), 130–143. 6

[LLZO02] LI T., LI Q., ZHU S., OGIHARA M.: A survey on wavelet applications in data mining. *SIGKDD Exploration 4* (2002), 49–68. 9

[MJL10] MOTTONEN J., JACOBS D., LIVESAY D.: Allosteric response if both conserved and variable across three chey orthologs. *Biophysical Journal 99*, 7 (2010), 2245–54. 2, 3, 4

[MWBF98] MILLER N., WONG P. C., BREWSTER M., FOOTE H.: Topic islandsÙa wavelet-based text visualization system. *Proceedings of the 9th IEEE Conference on Visualization* (1998), 189–196. 5

[MXJL09] MOTTONEN J., XU M., JACOBS D., LIVESAY D.: Unifying mechanical and thermodynamic descriptions across the thioredoxin protein family. *PROTEINS: Structure, Function, and Bioinformatics 75*, 3 (2009), 610–627. 4, 6

[NHL07] NIGHAM A., HSU D., LATOMBE J.-C.: Characterizing protein conformation space. *Singapore and MIT Alliance (SMA) Symposium* (2007). 2

[PKP10] PATIL A., KOLHE S., PATIL P.: 2d face recognition techniques: A survey. *International Journal of Machine Intelligence. 2* (2010), 74–83. 5

[PSW*06] POTZSCH S., SCHEUERMANN G., WOLFINGER M., FLAMM C., STADLER P.: Visualization of lattice-based protein folding simulations. *Proceedings of the Information Visualization* (2006), 89–94. 2

[QPA08] QAZI S., PANETTA K., AGAIAN S.: Detection and comparison of color edges via median based pca. *IEEE International Conference on Systems, Man and Cybernetics, 2008.* (2008), 702–706. 5

[Sal04] SALDANHA A.: Java treeview-extensible visualization. *Bioinformatics. 20*, 17 (2004), 3246–3248. 5

[SM07] SHEN Z., MA K.: Path visualization for adjacency matrices. *Eurographics/IEEE-VGTC Symposium on Visualization.* (2007), 83–90. 5

[SS02] SEO J., SHNEIDERMAN. B.: Interactively exploring hierarchical clustering results. *Computer. 35* (2002), 80–86. 5

[Swe96] SWELDENS W.: The lifting scheme: A custom-design construction of biorthogonal wavelets. *Appl. Comput. Harmon. Anal 3*, 2 (1996), 186–200. 5

[TLKT09] TALBOT J., LEE B., KAPOOR A., TAN D.: Ensemblematrix: Interactive visualization to support machine learning with multiple classifiers. *CHI 2009.* (2009), 1283–1292. 5

[VRP09] VARSHNEY S. S., RAJPAL N., PURWAR R.: Comparative study of image segmentation techniques and object matching using segmentation. *International Conference on Methods and Models in Computer Science.* (2009), 1–6. 5

[Wat05] WATTENBERG J.: A note on space-filling visualizations and space-filling curves. *Proc. IEEE Symposium on Information Visualization.* (2005), 181–186. 7, 9

[WB96] WONG P. C., BERGERON R. D.: Multiresolution multidimensional wavelet brushing. *Proceedings of the 7th IEEE Conference on Visualization* (1996), 141–148. 5

[YHW*07] YANG J., HUBBALL D., WARD M., RUNDENSTEINER E., RIBARSKY W.: Value and relation display: Interactive visual exploration of large datasets with hundreds of dimensions. *IEEE Transactions on Visualization and Computer Graphics 13*, 3 (2007), 494–507. 7, 9

[YR01] YEUNG K., RUZZO W.: Principal component analysis for clustering gene expression data. *Bioinformatics.* (2001), 763–774. 5

[ZHT06] ZOU H., HASTIE T., TIBSHIRANI. R.: Sparse principal component analysis. *Journal of Computational and Graphical Statistics. 15* (2006), 265–286. 5