Heterogeneity-based Guidance for Exploring Multiscale Data in Systems Biology

Martin Luboschik* Carsten Maus† Hans-Jörg Schulz‡ Heidrun Schumann§ Adelinde Uhrmacher¶

University of Rostock, Germany

ABSTRACT

In systems biology, analyzing simulation trajectories at multiple scales is a common approach when subtle, detailed behavior and fundamental, overall behavior of a modeled system are to be investigated at the same time. A variety of multiscale visualization techniques provide solutions to handle and depict data at different scales. Yet the mere existence of multiple scales does not necessarily imply the existence of additional information on each of them: Data on a more fine-grained scale may not always yield new details, but instead reflect the already known data from more coarse-grained scales – just at a higher resolution. Nevertheless, to be sure of this, all scales have to be explored.

We address this issue by guiding the exploration of simulation trajectories according to information about the deviation of the data between subsequent scales. For this purpose, we apply different dissimilarity measures to the simulation data at subsequent scales to automatically discern heterogeneous regions that exhibit deviating behavior on more fine-grained scales. We mark these regions and display them alongside the actual data in a multiscale visualization. By doing so, our approach provides valuable visual cues on whether it is worthwhile to drill-down further into the multiscale data and if so, where additional information can be expected. Our approach is demonstrated by an exploratory walk-through of stochastic simulation results of a biochemical reaction network.

Index Terms: I.3.8 [Computing Methodologies]: Computer Graphics—Applications; I.6.6 [Computing Methodologies]: Simulation and Modeling—Simulation Output Analysis;

1 INTRODUCTION

Systems biology aims at understanding the complex mechanisms that underlie both normal and defective modes of biological systems. Besides diverse experimental wet-lab methodologies and similar to many other areas, computer-aided modeling and simulation has become one of the key techniques in this field of research. Large amounts of data may thereby be produced, often exhibiting different behavior on different temporal and/or spatial scales – for example, different short, medium, and long term behavior within the generated time series data.

Resulting from complex non-linear interactions among a model's components, already the simulation of rather small models may exhibit dynamic behavior on multiple scales. In particular, this is the case for stochastic simulations [21], which take the intrinsic noise within biological systems into account and lead to rapid small-scale fluctuations in the trajectories. Here, a thorough analysis needs to be carried out in order to see through the noise and find interesting behavior within the data space. Multiscale data analysis becomes even more important when simulating multilevel models [13], which describe the complex interplay between components at different hierarchical levels of a system and typically operate at different spatiotemporal scales. However, the integrated analysis of data at multiple scales is a challenging task. The challenge lies in the desire to simultaneously analyze global long term behavior and detailed short term behavior, while both can hardly be achieved at the same time.

Formally, we define a *scale S* as a tuple (G,X) consisting of the *grain G* and the *extent X* [1, pp.55-65]. The grain defines a resolution as the lower threshold of observation – everything that is smaller than the grain cannot be observed on this scale. This is complemented by the extent, which defines a value range as the upper threshold of observation – everything exceeding the extent cannot be observed either. For a visual analysis, the displayable number of grain-sized fragments in a given extent $X \div G$ is bounded by the screen resolution. If there are more of these fragments than available pixels, the trajectory cannot be represented faithfully on this scale. And even if it is possible to depict all data, the actually desired information may be hidden underneath it: Strongly fluctuating fine-grained data, such as noise, can superimpose lower scale behavior up to the point where it can no longer be recognized. This effect is known as *masking* [22]. For both problems – the multiscale challenge and the masking – exist a number of established solutions (cp., Sec. 2). In essence, most approaches display the data individually at different scales and link these displays via drill-down/roll-up to ensure their faithful depiction while still giving access to the broadest extent and finest grain.

Yet, none of the existing approaches gives consideration to the question if there even is noteworthy information to display at each scale. Often, data sets lack fine-grained behavior in certain parts, so that its interactive exploration through drill-down operations to the most fine-grained scale is not necessary. Whereas in other parts, subsequent drill-down operations will unveil new behavior all the way down to the finest grain. Current multiscale visualizations do not guide the user in this respect. Regardless of the data, they generally require an analyst to explore the full data set by always drilling down to the scale of highest resolution just to be sure that no more interesting data is hidden there. This is a time-consuming task, which we strive to improve in this paper.

According to this, we aim to support systems biologists in finding desired information at fine-grained scales without the need of an exhaustive search. For this purpose, we contribute an approach that interprets the heterogeneity of simulation results at subsequent scales as an indicator for noteworthy information. Regarding to this, it depicts visual cues where a drill-down to fine-grained scales may be valuable. As detailed in Sec. 3, our approach comprises multiple steps: A calculation step in the *data space*, which computes dissimilarities between subsequent scales and aggregates these values to regions of homogeneity/heterogeneity. The representation step in the *visual space* uses the gathered heterogeneity information to enhance time-course data visualization. For that, the heterogeneity with respect to subsequent scales is visualized alongside the data. The adjustment step in the *user interface space* (UI

*e-mail: luboschik@informatik.uni-rostock.de
†e-mail: carsten.maus@uni-rostock.de
‡e-mail: hjschulz@informatik.uni-rostock.de
§e-mail: schumann@informatik.uni-rostock.de
¶e-mail: lin@informatik.uni-rostock.de

space) allows for an informed drill-down to adjust the scale of the visualization by providing different interaction facilities.

This novel approach is exemplified with a use case that deals with the exploration of simulation trajectories from a biochemical reaction network describing the control of the cell cycle in yeast. This exemplary data was presented to application experts who used our approach to investigate it at multiple temporal scales. We report on their insights in Sec. 4 and summarize their feedback and hurdles to adoption in Sec. 5. Lastly, current and future research directions for the presented approach are outlined in Sec. 6.

2 Related Work

In the current literature, multiscale data is handled in the data space, in the visual space, as well as in the space of the user interface.

In **data space**, different techniques are used to first extract the data on multiple scales. The most common approach is a simple binning into scales of predefined grain – e.g., the duration of a cell cycle, or time intervals consistent with experimental data acquisition, such that a comparison with wet-lab results can be made. This method has the convenient side effect of aligning non-equidistant time steps, as they occur, for example, in stochastic simulations [14]. If no semantically meaningful scales exist, abstraction techniques, such as generalization or aggregation, are used subsequently to create an ordered set of scales from a given data set [2]. A third possible way of isolating data at specific scales is the application of filters to the data (e.g., low-pass filters), which is often used in signal-processing applications. Some recent approaches in systems biology go even further by actually determining which of the modeled and subsequently simulated processes exhibit behavior on which scale and then reduce the model to gain only data on a scale of interest [17]. In general, it can be said that these methods are useful to counter the masking problem, as the extraction of individual scales breaks down the overall system behavior into the long-term and short-term aspects of which it is composed.

Once the scales are separated, they are brought back together in **visual space** to nevertheless allow for their integrated exploration. For this, a number of different options are used, which basically follow the design space of such visual combinations [10]. The simplest way is the use of two juxtaposed or integrated views (depending on their linking) [9, 23]. The most complex way is to superimpose the two views in such a way, that the masking effect is not reintroduced [2, Fig.1]. Yet, a recent study [8] shows that these methods of composition are less effective than more advanced overlay compositions, which use a piecewise embedding as a compromise between these two options [5]. In addition, methods using nesting are also known, which insert depictions of local, fine-grained features (e.g., outliers or peaks) in a global, coarse-grained visualization [6, 15]. These methods have in common, that they aim to provide a coarse-grained overview together with a fine-grained detail view through different means of integration – thereby providing solutions to the multiscale challenge.

To manipulate such visual compositions of multiple scales in order to change the displayed extent and grain, a range of possible mechanisms can be utilized in the **UI space**. Standard methods, such as drill-down and roll-up [2] permit to move back and forth between different scales, uncovering additional details or gaining a broader overview putting the details into context, respectively. For mixed-scale visualizations using nesting or overlay composition, interactive lens techniques [11, 24] can be used to embed and manipulate the client visualization on one scale inside or on top of the host visualization of a different scale. The interaction on mixed-scale visualizations can be thought of as an unbalanced drill-down/roll-up, where the scale of a view is not adjusted globally, but locally. Again, these methods serve to interactively find and fine-tune a suitable tradeoff between local details and global overview to resolve the multiscale challenge.

All of these approaches permit to visually access and explore multiscale data. Yet, none of these techniques asks the question whether it is actually necessary to explore all scales of that data. This is an important question, as the straightforward extraction, visualization, and interactive manipulation of scales does not necessarily mean that there exists new information in each of them. A drill-down may or may not reveal unexpected details, which are smaller than the current scale's grain and therefore hidden just below the lower threshold of observation. A roll-up on the other hand may or may not yield a discovery of a global trend, which is larger than the current scale's extent and therefore hidden just outside the upper threshold of observation. So, in order to use the established visualization and interaction methods for multiscale data in an informed way, a user must be made aware of possibly interesting aspects on different scales and thus be guided towards them. The approach introduced in the following section addresses this issue.

3 An Approach for Multiscale Guidance

In order to eliminate the necessity of extensively exploring the data at all scales, the main goal of our work is to provide interscale and intrascale guidance for systems biologists analyzing their simulation trajectories. This means to provide visual cues at which scale (interscale guidance towards a specific grain) and in which region of this scale (intrascale guidance towards a specific extent) the trajectory shows deviating and therefore possibly interesting behavior. To achieve this, we depict where a more fine-grained scale brings in additional information as compared to more coarse-grained scales. This indicates potential targets for a further drill-down, but it also indicates which coarse-grained scale may suffice for communicating certain behavior of the trajectory. For this purpose, we examine subsequent scales with respect to heterogeneity between their data. We interpret heterogeneity between scales as an indicator for noteworthy additional information at a more fine-grained scale.

Our proposed approach comprises of the following steps, which are discussed in more detail in the following sections:

I. Computing the heterogeneity values between subsequent scales of a given multiscale data set.

II. Visualizing the heterogeneity values in addition to the data on a currently selected scale.

III. Interacting with the heterogeneity visualization to perform a guided navigation across and within scales.

3.1 Computing the Heterogeneity Values

This first step does not directly compute the concrete grains and extents towards which to guide the user, but instead uses an indirect approach of computing the heterogeneity between scales as indicator where an in-depth analysis could be worthwhile. As input, this step assumes a data set with scales S_i $(1 \leq i \leq n, n \in \mathbb{N}, n \geq 2)$ ordered from coarsest (S_1) to finest grain (S_n). These scales are either inherently given by the data set or extracted through methods, such as those mentioned in Sec. 2. The computation is then performed between every pair of subsequent scales S_i and S_{i+1} and consists of the following 3 steps:

1. Determination of specific points in data space where to calculate the heterogeneity.

2. Calculation of the heterogeneity according to a given metric.

3. Aggregation of the point-wise computed heterogeneity into intervals.

The last step can be seen as the inverse to the first step: The heterogeneity computed at specifically chosen points in data space is

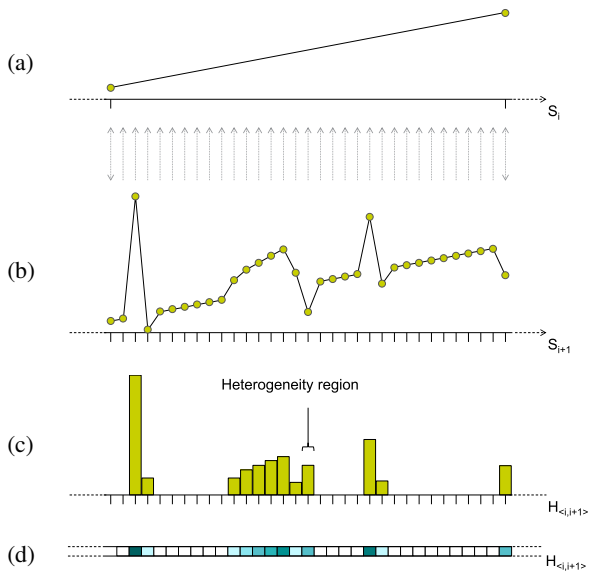


Figure 1: To determine heterogeneities between subsequent scales S_i and S_{i+1} we simply map the data points of one scale at the respective other scale (a, b). The calculated heterogeneity values are aggregated into heterogeneity regions of principally arbitrary size (c). To provide intrascale guidance, we choose the grain of S_{i+1} as the regions' size. A compact representation is gained by so called *heterogeneity bands* mapping heterogeneities to color (d).

turned back into a continuously defined step function covering the whole data space. While in theory, it would suffice to aggregate the heterogeneity values between two subsequent scales into a single value describing their overall discrepancy, this would only aid interscale guidance – giving information on whether or not further drill-down is necessary. For intrascale guidance, it is the splitting up of the aggregation over multiple intervals or regions, which lets the user also pinpoint where on a given scale the drill-down will actually yield additional information and where not.

Determining Data Points At first, we determine specific points in data space, called *calculation points*, at which the heterogeneity is to be computed. This can be done in multiple ways. For instance, each data point given at particular scales can be chosen as a calculation point. Yet at more fine-grained scales, this may result in a large number of such points. Alternatively, sampling (e.g., taking every n^{th} data point) or approximation (e.g., taking characteristic points, such as inflection points or saddle points) can be used to reduce the number of calculation points. While approximation captures the characteristics of the data more closely, it has the drawback of creating non-equidistant intervals based on the data's features [25]. Sampling on the other hand may miss crucial data features. Thus for different models and different types of simulations, this decision has to be made by the user to choose an approach that fits the data at hand and its generating process. Without loss of generality and for the sake of clarity, we use every given data point as a calculation point in the course of this conceptual discussion, although it may be computationally expensive in practice.

It should be noted that on two subsequent scales, the calculation points on the fine-grained scale are generally not aligned with those on the coarse-grained scale. To achieve such an alignment, we use a straightforward approach that maps the calculation points of both scales onto the respective other scale (Fig. 1a,b). The data values at these mapped points are determined through an interpolation approach (linear, cubic, nearest-neighbor…), again depending on the type and general behavior of the simulation.

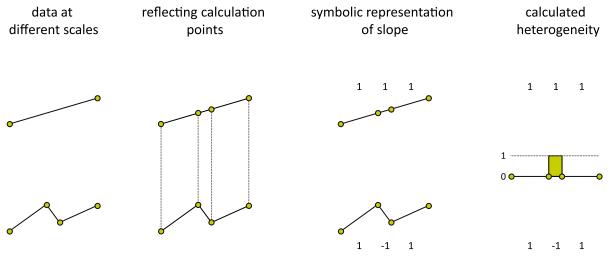


Figure 2: Our metric detects changes in the slope between subsequent scales. If slopes of subsequent scales differ in their direction, a heterogeneity value of 1 is assigned – otherwise 0.

Computing Heterogeneity Values Given the calculation points and a faithful interpolation method, the heterogeneity is computed in this second step. To actually quantify the heterogeneity between two calculation points, different metrics can be used. Which metric to use depends on the current task of the systems biologist and the application scenario. For example, if absolute data values are of concern, the metric should be sensitive to value differences (e.g., absolute differences in Euclidian space). Yet, if trends are in focus of the examination, differences between the trends of subsequent scales should be captured (e.g., SAX-distance [16]). More sophisticated metrics that detect differences in patterns may also be useful. Beside these established metrics, we developed a novel metric that is inspired by [12]. For it, the slope of one calculation point to the next is represented symbolically by its sign: increase $= 1$, steady $= 0$, decrease $= -1$. If the sign of the slope differs for the examined subsequent scales, a heterogeneity value of 1 is assigned, otherwise the heterogeneity is set to 0 (see Fig. 2). While being computationally inexpensive, this metric captures just enough information to be used for trend analysis for stochastic event-based simulation trajectories. It is an example of how metrics targeted towards certain applications can be used interchangeably at this step. Consequently, we provide an extensible set of different metrics to the systems biologist – with each of them potentially providing cues for an exploration at more fine-grained scales.

Aggregating Heterogeneity Regions While a single aggregated heterogeneity value cannot guide the user to specific extents on a scale, the set of all point-wise defined heterogeneity values may exceed the available screen resolution denying their faithful representation. Hence, local *heterogeneity regions* are established (Fig. 1c), which permit a better localization than a single value, but prevent an overplotting and thus misperception of too many values. Heterogeneity regions are intervals in data space for each pair of subsequent scales, which accumulate the heterogeneity values. As a suitable region size, we chose the minimum distance of data points at the more fine-grained of the two subsequent scales. This way, we assign at least one calculation point and thus at least one heterogeneity value to each region. To handle the resolution problem of very fine-grained scales, a region size corresponding to a minimum screen size (e.g., 2 pixels) can be used as a lower bound.

Once the regions are established, their heterogeneity values are determined by aggregating the individual heterogeneity values from each calculation point in a region. This aggregation has to preserve the necessity to explore a heterogeneity region if one of the included calculation points expressed this through a high heterogeneity value. Therefore, we generally determine the value of a heterogeneity region as the maximum heterogeneity over all calculation points. If the heterogeneity values express only the existence or absence of deviations between scales, the weighted average can be used to indicate how much additional information is likely to appear – for example, when utilizing our simplified metric.

It is noteworthy that the calculation of heterogeneity values and their aggregation to heterogeneity regions is freely adaptable. Each aspect for calculating heterogeneity values can be adapted or exchanged to tailor the overall approach to different analysis goals from different applications. Therefore, we provide a set of different metrics, as well as interpolation and aggregation methods, which are adjusted to the need of systems biologists. Thus, their interactive adaptation becomes a part of the overall exploration process. At this point – regardless of the chosen approach – we have extracted a set of heterogeneity regions $H_{\langle i,i+1\rangle}$ calculated for each pair of subsequent scales S_i and S_{i+1}. These are to be visualized in the next step, to communicate them to the user to aid an informed analysis.

3.2 Visualizing the Heterogeneity Values

This second step deals with the visual representation of the heterogeneity regions in conjunction with the actual data. In general, we map the heterogeneity regions and their values for each scale to small colored rectangles and align them to form a so-called *heterogeneity band* (Fig. 1d). These bands are stacked on top of each other in order of their granularity (Fig. 3). In the following, we discuss different aspects of the heterogeneity visualization, such as the chosen colors, the number of heterogeneity bands, the visualization along the multiscale data, and the extensibility of our approach.

Utilizing Different Color Scales In principal, the used color scale can be freely chosen, but it should clearly separate heterogeneous and homogeneous regions. To achieve this, we use a color scale ranging from white for regions exhibiting no heterogeneity to a second saturated color for high heterogeneity. This way, highly saturated regions stick out naturally as prime targets for further drill-down. The color scale may be applied to the local heterogeneity range of each individual heterogeneity band or to the global heterogeneity range of all bands. Through choosing a local or global color scale, the analysis can be guided either qualitatively (local range: "where" is additional information) or quantitatively (global range: "how much" additional information).

Reducing the Number of Heterogeneity Bands The heterogeneity visualization begs the question, whether always all bands have to be shown. The upper limit of theoretically necessary bands to capture heterogeneities between all scales lies at $log_2(g)$ with g being the finest grain. This is due to the fact, that at least a bisection of a given grain is needed to ensure possible new behavior in each interval on each scale. So, for an assumed 1D simulation trajectory of $1,000,000$ data points at the scale of highest resolution, showing about 20 bands simultaneously ensures the theoretical visibility of additional information at every scale and region (since $2^{20} > 1,000,000$). Yet, oftentimes some subsequent heterogeneity bands do not differ very much and can thus be collapsed into a single band to save screen real estate. For this, we compute the similarity of subsequent bands by summing the squared heterogeneity differences per region and applying a simple, interactively adjustable threshold to them.

Combining Heterogeneity and Data Visualization Heterogeneity bands alone do not show the actual data. Thus, they must be combined with the data visualization, for which we employ two commonly used principles: superimposition and juxtaposition.

In the superimposed mode, the heterogeneity bands are enlarged to fill the available space and the trajectory is plotted on top of them for a selected scale of interest (Fig. 3a). Superimposition works only if the data can be visualized by a sparse representation (e.g., line plot, scatterplot, etc.), which leaves enough whitespace for the heterogeneity information to "shine through" and does not interfere with the used color-coding. In this representational mode, the user can match the heterogeneity regions in the background with the data

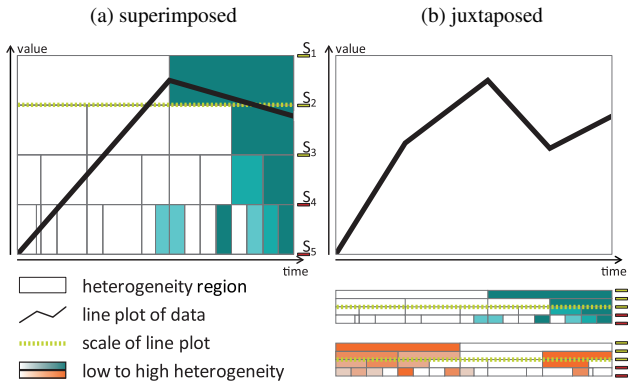


Figure 3: The principle design of our visualization approach: Heterogeneity bands are stacked and visualized either behind (a) or next to the data (b). In both cases, saturated colors indicate changes between the data of subsequent scales. The green dashed horizontal cursor indicates the scale of the currently visualized trajectory and the small colored bars to the right show, if the data plot of a scale is affected by overplotting (red) or not (green).

visualization and identify parts of the plot that may reveal additional information on other scales than the currently shown one.

If superimposition is not possible, the views can always be juxtaposed. In this case, the heterogeneity bands are placed in a compact form below the data visualization (Fig. 3b). By doing so, the heterogeneity bands serve as an annotation to the data display. The user can switch between both modes upon demand.

Both displays feature the same set of additional indicators. A cursor (green dashed line in Fig. 3) points to the currently selected scale and its related heterogeneity bands: The heterogeneity band above the cursor depicts the deviation to the next more coarse-grained scale, the band below the cursor shows the deviation to the next more fine-grained scale. The currently selected scale may be too fine-grained and thus exceed the current screen resolution, resulting in overplotting. In the spirit of the visual uncertainty display from [7], it is communicated by red indicators next to the bands for which scale overplotting and thus an information loss occurs. On the other hand, green indicators signal that a scale is not affected by the resolution problem. These indicators provide additional cues for guiding the analysis. For example, to resolve the resolution problem at a region of interest, further zoom-in reduces the extent of the currently shown part of the trajectory and thus also the overplotting.

Extending the Visualization Up to this point, our considerations concern multiple scales of a single dimension. Yet, the presented approach is applicable to further dimensions in the very same manner. Additional data dimensions can be introduced either through multivariate data or through genuine 2-dimensional or even higher dimensional simulation data, as it is produced, for example, by cellular automata simulations [18]. In essence, this is addressed by computing the heterogeneity for each dimension individually, yet jointly displaying them to allow for a back and forth, and a comparison between them. The heterogeneity bands of different dimensions may guide towards different features in the data. This extension to further dimensions is illustrated in Sec. 5.

The same approach is used to compute and visualize multiple heterogeneity metrics at once. As each metric captures specific data characteristics, it is hard to pick *the best* metric for an unknown data set and maybe even without a clear analysis goal, yet. By showing multiple heterogeneity bands of different metrics (Fig. 3b), the user can, for example, explore value-based and trend-based metrics and their respective features in the data simultaneously.

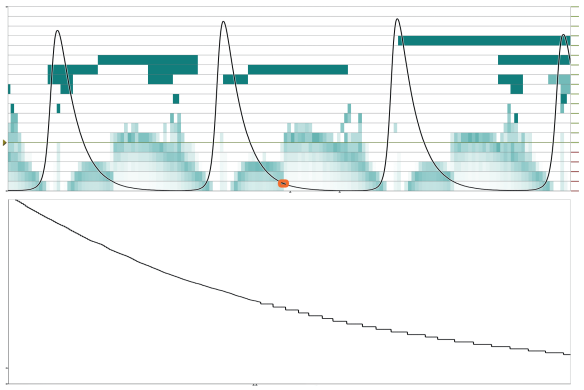
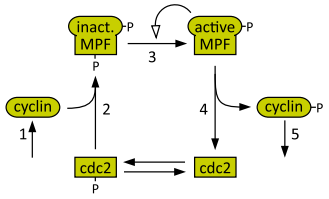


Figure 4: Within a simulation time series, a repeating pattern at the more fine-grained scales attracts attention (a) and unveils a faulty switch within the simulator's behavior (orange box enlarged in b).



$$\xrightarrow{k_1} cyclin \quad (1)$$

$$cyclin + cdc2 \xrightarrow{k_2} MPF_i \quad (2)$$

$$MPF_i \xrightarrow{k_3} MPF_a \quad (3)$$

$$MPF_a \xrightarrow{k_4} cyclin\text{-}P + cdc2 \quad (4)$$

$$cyclin\text{-}P \xrightarrow{k_5} \quad (5)$$

Figure 5: Reaction network graph and corresponding biochemical reaction equations of the example model. (De-)phosphorylation of cdc2 is neglected. Parameters used for simulation: $k_1 = 0.015 cdc2_{tot}\, min^{-1}$, $k_2 = 200 min^{-1}$, $k_3 = k'_{act} + k_{act} MPF_a / cdc2^2_{tot}\, min^{-1}$, $k_4 = k_d / V\, min^{-1}$, $k_5 = 0.6 min^{-1}$ $cdc2_{tot} = 10^5$, $k'_{act} = 0.018 min^{-1}$, $k_{act} = 180 min^{-1}$, $k_d = \{2,3\} min^{-1}$ (depending on simulation experiment), the volume V is a relative measure, which is either fixed at 1 or increases exponentially from 1 to 2 (within a time range of $116 min$) and is immediately reset to 1 afterwards to mimic cell division.

3.3 Interacting with the Heterogeneity Visualization

The first two steps discussed in Sec. 3.1 and 3.2 compute and communicate regions on the different scales towards which to guide a user. This last step is the one, which actually enables the user to get to these regions and inspect the corresponding part of the trajectory at any chosen scale. Conceptually, two mechanisms have to be provided to do so: one to follow the interscale guidance to a specific scale and one to follow the intrascale guidance to a specific region on that scale. For the first of the two, the back and forth between scales, commonly used mechanisms are drill-down and roll-up. For the second of the two, the steering of the exploration towards a specific region, a zooming-in and zooming-out into these regions, as well as a panning from one region to another are simple ways to support the navigation on a given scale. Zooming has the problem, that it reduces the shown extent in order to depict the region of interest in more detail. If this is not desired and needs to be prevented to keep the overview of the entire extent, more sophisticated approaches, such as a lens with fisheye distortion can be employed. For both – zoom and distortion – it is important to adapt the heterogeneity bands besides or underneath the data plot accordingly, so that both remain aligned (see Fig. 7d in Sec. 4). This applies also to the overplotting indicators, which are also adapted instantly. Likewise, the mentioned green cursor in between the heterogeneity bands is adjusted according to the drill-down and roll-up operations.

For a more direct access to a region of interest on a scale of interest, this region can also be selected directly from the heterogeneity bands. This combines interscale and intrascale navigation into a single interaction to quickly leap from one region on one scale of interest to another region on another scale of interest.

In addition to these fundamental interaction methods to reach all parts of the data, interaction should be provided, which allows for (re-)parametrizing the calculation and visualization steps. These simply provide interactive handles for the multiple computational and representational decisions mentioned in the previous steps: For the computation step, this includes different choices for metrics to compute, for interpolation methods to use, etc. For the visualization step, this involves the switching between global and local color scales, as well as between juxtaposed and superimposed composition of the trajectory display and the heterogeneity bands.

The following section will illustrate how these three steps and in particular the different interaction methods provide the necessary guidance and the means to follow the guidance.

4 Application Scenario

Stochastic simulations are of considerable importance for systems biology. They allow for taking the intrinsic noise into account, which originates from small variances in the speed of biochemical reactions, particularly in the case of low copy numbers of involved molecules. Large data sets may thereby be generated, which typically show different short and long term behavior on multiple temporal scales and thus need visual support for their explorative analysis. In this section, we demonstrate results of our approach in this domain. We start with a motivating example, before we give a brief explanation of the simulated model from systems biology and subsequently illustrate the guidance through multiscale trajectories.

4.1 Motivating Example

We start with an example showing the trajectory of a Lotka-Volterra system [20] describing a predator-prey relation (Fig. 4). Our novel metric reveals a repeating pattern at the most fine-grained scales that attracts attention due to its uniformity. A drill-down and zoom-in to one of those regions exposes a regular staircase behavior (Fig. 4b), which was identified as a faulty switch during the simulator's execution. This way, our approach directed the user towards additional information, which was invisible at the coarse-grained scales. Thereby our approach helped to discover and fix the previously unknown bug within the used simulation tool. This motivated us to take our approach to more complex biological models and facilitate guidance during their exploration.

4.2 Use Case

Our example model is based on a reaction network by Tyson [19] describing a simple biochemical control circuit of the cell division cycle in fission yeast cells. Fig. 5 provides an illustration of the network and the corresponding biochemical reaction equations. In short, the model consists of two proteins: cyclin and cdc2. Both can form the so called MPF complex, which may be either in an inactive state (MPF_i) or in an activated state (MPF_a). MPF activation, i.e., dephosphorylation of the cdc2 subunit, is described by an autocatalytic process. That means, the more MPF_a exists, the faster the process of activation of further MPF will be. Tyson identified ranges of certain model parameters where regular oscillations with bursts of the amounts of inactive and activated MPF complexes can be observed. In addition, he used the model to analyze the effect of cell growth on the dynamic behavior by assuming an exponentially decreasing rate coefficient (k_4) of the MPF dissociation process.

Simulating this model generates data describing the dynamic change of molecule amounts over time. Unlike Tyson, who used numerical integrations of deterministic ordinary differential equations, we simulated the model in a stochastic event-based manner by applying the Gillespie algorithm [4]. As a result of this kind of simulation, the data easily exceeds a multitude of hundreds of thousands or even millions of data points and thus does not allow for visualizing the data in every detail in the limited screen space. Therefore, it is necessary to analyze the data on multiple scales.

The data comprises one single fine-grained scale containing every event of the simulation run. Other scales are merely implicitly contained. Yet it may already suffice to look for and describe certain behavior at a coarser grain, while at the same time coarser scales also reduce the risk of losing important information due to masking. Hence, we transformed the data into a multiscale data set by applying low-pass filtering on the simulation data (see Sec. 2). This iteratively reduces the grain at each scale by a factor of 2 and uses the average value of subsequent data points for the coarser scales.

4.3 Initial Setup

The initial configuration of the exploration relates to the first two steps of our approach: the calculation of heterogeneities (Step I) and their visualization (Step II). In Step I, every data point is also used as a calculation point to prevent any loss of information, and a linear interpolation is used for their alignment between scales. The choice of the latter is motivated by the fact that all of our data is increasing or decreasing monotonously with no gaps or sudden jumps that would prevent a linear interpolation. As a suitable metric, we chose our novel metric as we are mainly interested in trends and thus in changes of the slope. However by detecting changes of the slope, this metric also detects marginal changes, such as noise. Therefore, we additionally apply metrics that quantify the change of the slope (delta of the gradient) and that are sensitive to value changes (absolute differences). Lastly, we chose an averaging of heterogeneity regions for the aggregation of qualitative metrics and a maximum aggregation for quantitative ones.

For Step II, we use the juxtaposition mode for annotating the data display with the heterogeneity bands. This is due to the fact that we provide heterogeneity information based on multiple metrics. The data of the currently selected scale is presented by line plots, as they are commonly used in systems biology to visualize simulation traces. The initial scale shown in the plot is the finest scale at which the data visualization is not affected by overplotting.

4.4 Guided Data Exploration

In several sessions, we explored the simulation data in close cooperation with five systems biologists. We summarize these sessions in the following to illustrate the different exploration courses we observed. They highlight the usefulness of our heterogeneity visualization for interscale and intrascale guidance. Accordingly, these sections relate to the interactive exploration Step III of our approach applied to the use case example. The simulation results visualized in the following figures comprise roughly 3.6 million data points.

4.4.1 Interscale Guidance

To start the exploration process with a maximum number of data points right from the beginning, the initial view is set to the finest scale at which the visualization is unaffected by overplotting. This also prevents from depicting oversimplified behavior, which is responsible for large heterogeneity values at the most coarse-grained scales. In our use case, we thus typically select one of the mid-range scales to start with, from which a roll-up presents a more abstract view. Conversely, to gain a deeper understanding of the more detailed, local aspects of the simulated model, a drill-down from this initial scale to more fine-grained scales is needed. Without guidance to noteworthy points in the data, this becomes a time-consuming task, which might not even reveal further information.

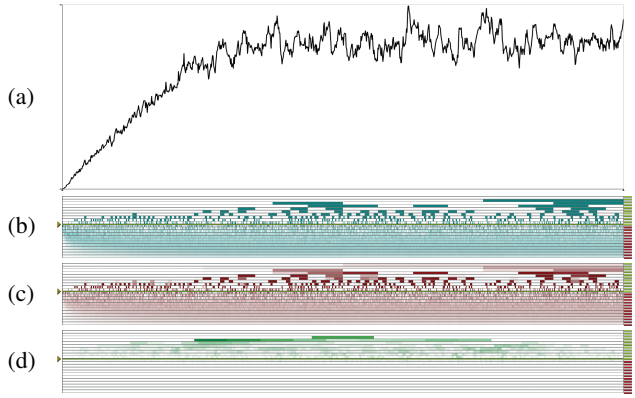


Figure 6: Trajectory of MPF$_a$ from a simulation with $k_4 = 2\,min^{-1}$, i.e., $k_d = 2\,min^{-1}$ and the cell volume $V = 1$ is constant (a). Our metric reveals equally distributed changes in slope throughout the whole trajectory (b) suggesting the existence of noise. A second metric quantifying slope changes bares the same behavior (c). A third metric (d), which is sensitive to value changes, reveals only marginal changes at more fine-grained scales.

For example, in a simulation experiment with a constant dissociation rate coefficient $k_4 = 2\,min^{-1}$, i.e., without taking the growing cell volume into account, the amount of active MPF increases linearly at the beginning and reaches afterwards some kind of noisy steady state (Fig. 6a). Since the current scale of the view is an intermediate one, further drill-down to investigate this behavior is possible. Yet at the same time, the overplotting indicator at the side of the multiscale view in Fig. 6b shows the user that a drill-down to more fine-grained scales would also hide some information due to the limited screen resolution. While this can be alleviated by additionally zooming-in, the question is, whether there is worthwhile information to be found there to justify taking all these steps to see the data on a more fine-grained scale? The heterogeneity bands of our novel metric reveal many slope changes at more fine-grained scales throughout the whole trajectory (Fig. 6b). However, these changes seem so equally distributed across the entire trajectory and their number decreases at more fine-grained scales, suggesting that at these scales solely noise can be observed. To quantify the slope changes, we subsequently apply a slope-sensitive metric, which measures the change of the gradients and maps it linearly to the interval $[0 \dots 1]$, where 0 stands for having the same gradient and 1 stands for having opposite gradients by 180°. The corresponding heterogeneity bands bare nearly the same behavior as our metric did (Fig. 6c) and thus underscore the first hypothesis. Finally, the absolute difference metric reveals that all changes detected by the first two metrics are of a very small amplitude and nearly no changes appear at the more fine-grained scales (Fig. 6d). Moreover, this metric shows that all major changes appear only at coarse-grained scales. Therefore, beyond a single drill-down into an arbitrary region to observe and confirm the hypothesis of noise, no further examination needs to be done and the initially selected scale suffices for data representation. However, a roll-up to more coarse-grained scales is useful to get an overview of the longer term behavior of the trajectory, which is undisturbed by the noise found on the medium scales.

Taken together, the explored simulation results suggest that stochastic perturbations are too low to transiently activate MPF (cf. [19, Fig. 3b]) and thus the stochastic model also needs to take the dynamic volume change into account to capture the system's behavior of regular oscillations.

4.4.2 Intrascale Guidance

Besides pointing the user to a scale of potential interest, it is also crucial to guide to interesting areas *within* a given scale to reduce the effort of finding desired information in this regard as well.

For example, the amount of MPF_a shows a regular, oscillating pattern in a simulation experiment comprising a volume-dependent rate coefficient k_4 (Fig. 7a). Our metric reveals an uneven distribution of slope changes at fine-grained scales, which suggests the presence of noise again. Although noise detection is generally reduced in the regions of steep peaks (indicated by gaps in the heterogeneity bands), the metric indicates heterogeneities at the most fine-grained scales in those regions (orange arrows in Fig. 7a). As this may be due to noise, as it would be normally detected, we additionally apply the absolute difference metric in combination with a local color scale to get more indications for interesting behavior (Fig. 7b). This metric shows that the corresponding changes clearly differ from the surrounding ones and therefore initiates the exploration of those peaks in more detail. Without guidance, this would require to zoom-in to the trajectory and to navigate towards the region of interest by repeatedly panning the visualized data section. Hence, interaction techniques in terms of interscale and intrascale guidance are used to navigate to those regions more directly.

The first steps are to drill-down to the most fine-grained scale and to expand the visualized region at a peak by switching to the fisheye view (Fig. 7c). As still no noteworthy behavior can be spotted due to strong overplotting, the next step is to select the region of interest, such that the extent of the trajectory is reduced and thereby the fluctuating behavior at the top of the peak becomes visible (Fig. 7d). Without any indication that it is there, this behavior is likely to be overlooked as it is only visible at the most fine-grained scales.

The behavior of rapid small-scale fluctuations found within the exploration process can be explained by the strong competition between activating (Reaction 3) and dissociation events (Reaction 4) at the turning point of the peaks. The steeply increasing amount of MPF_a activates more and more MPF complexes, but at the same time also increases the speed of the dissociation reaction. Finally, the dissociation overturns the activation (i.e., by becoming faster than it), which leads to a rapidly decreasing amount of active MPF. In this way, our approach pointed to noteworthy information at fine-grained scales, that finally suggest a correct working stochastic simulation faithfully reproducing the reaction network behavior.

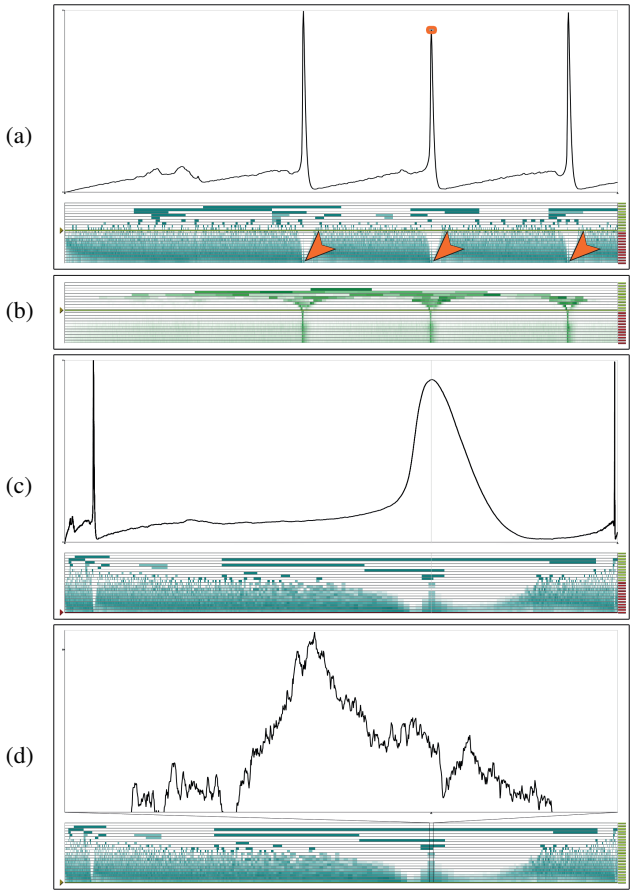


Figure 7: Trajectory of MPF_a from a simulation with $k_d = 3$ and dynamic cell volume (a). Our slope-based metric reveals an uneven distribution of heterogeneities for finer grains with heterogeneities at the peaks (arrows). A metric sensitive to value changes reveals a clear differentiation from surrounding areas at those scales (b). Applying a fisheye lens preserves a distorted overview (c). Only a drill-down to the finest grain and following magnification at those areas reveal fluctuation at the top of a peak (d).

5 Discussion

When reflecting on the conceptual limitations and evaluating our approach with system biologists, its generality and the interchangeability of the employed computational methods turn out to be the most important aspects to address.

In case of the conceptual limitations, noise poses a problem to our approach, which is inherent in its design: Noise could either be falsely detected as a noteworthy heterogeneity information and obfuscate the actual regions of heterogeneity, or actual regions of interest could falsely be disregarded as noise. This is a known problem of its own, which is not specific to our approach. Yet, it nevertheless has to be addressed by us to aid the user in dealing with data having unknown noise characteristics, as it would be the case in an exploratory scenario. Here, the interchangeability comes into play, as it not only permits for exchanging heterogeneity metrics for alternatives, but also to use and display multiple such metrics in conjunction. Since different metrics are sensitive to different characteristics of a signal and thus prone to pick up different kinds of noise, their joint use for the exploration alleviates this problem already to a large degree. Fig. 7a and 7b are examples, how different metrics help to distinguish noise from noteworthy information.

Interestingly, it was exactly this generality that makes for a large part of the capabilities and potentials of our approach, which sparked most discussions with the systems biologists. The variety of different metrics and the understanding of their individual effects took a while to communicate. Yet, the learning effort was lowered by the fact that every application partner could easily load his or her own data from simulations they conducted themselves. The familiarity with the shown data eased them into the approach, as they could experiment with the different metrics and see first hand, which known aspects of the data they picked up on and which not. Using the extension for two-dimensional data (cf. Sec. 3.2), we were even able to accommodate the needs of those colleagues who work with cellular automata simulations to examine spatial phenomena. An example of such a two-dimensional simulation trace is given in Fig. 8. Quick access to the different metrics via keyboard shortcuts (e.g., next metric, previous metric), as well as the overplotting indicator next to the heterogeneity bands were introduced following the suggestions from our application partners.

The overall feedback from our partners from the systems biology domain, as well as from collocated medical experts and engineers working with multiscale data on the same project, was outright positive. Our approach fills a clear gap in the current state-of-the-art of simulation software, which usually provide a mere line plot for trajectory visualization – if at all. Our software will be made available

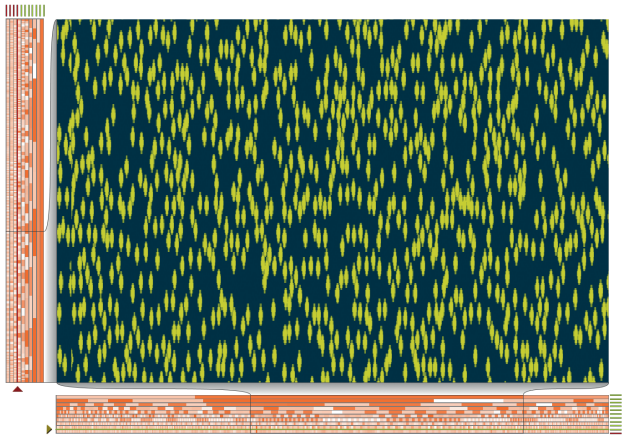


Figure 8: Our approach applied to 2D multiscale data gained from cellular automata simulation of *lipid rafts* at the cell surface.

for the JAMES II simulation framework [3]. This integration with a plug-in-based simulation software, which allows for replacing simulation engines, will enable users to re-simulate regions of interest. This way, a more fine-grained view can be gained locally by a more precise, yet also more time-consuming simulation run on demand.

6 Conclusion and Future Work

In this paper, we presented an approach to guide the analysis of multiscale data to regions, where the behavior of the data deviates between subsequent scales. These regions are captured through heterogeneity metrics, added as overlay or annotation to the existing data visualization, and used for a guided and thus less time-consuming exploration of the data. This approach was motivated by and devised in collaboration with domain experts from systems biology, where large quantities of data are generated and require efficient ways for their analysis. Through its rich set of parameters, it spans a broad range of applications in the field of systems biology.

Currently, we are exploring this range in particular for data from high-throughput wet-lab experiments, such as high-resolution microscopy images and clustered heatmaps of gene expression data. While our approach was developed with the commonly used representations from systems biology in mind, it is general enough to envision its application to other representations in a similar vein. In this regard, we are currently adapting it to higher dimensional data, as well as to different multivariate visualization techniques, such as parallel coordinates, adjacency matrices, and treemaps.

As different data from different domains lead to different analysis questions, other – possibly novel – heterogeneity metrics, approximations, and interpolations have to be investigated in future work to faithfully capture the desired differences between scales. For example, extremely noisy data requires metrics, which take the results of a preceding noise analysis into account, so that even for data with a low signal-to-noise ratio, distinct exploration targets can be extracted. Yet, additional metrics also increase the burden on a user to choose those applicable to the given data and suitable to capture the behavior the user looks for. To aid the user in choosing a fitting metric, enhancing the visual interface with statistical information promises to be a valuable first step in this regard.

Acknowledgements

This work was funded by the German Research Foundation (DFG). The authors thank the Modeling&Simulation group of the University of Rostock for their valuable feedback, as well as Clemens Holzhüter for inspiring discussions and Steffen Hadlak for his help with the implementation.

References

[1] V. Ahl and T. Allen. *Hierarchy Theory: A Vision, Vocabulary, and Epistomology*. Columbia University Press, 1996.

[2] N. Elmqvist and J.-D. Fekete. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE TVCG*, 16(3):439–454, 2010.

[3] R. Ewald, J. Himmelspach, M. Jeschke, S. Leye, and A. M. Uhrmacher. Flexible experimentation in the modeling and simulation framework JAMES II – implications for computational systems biology. *Brief. Bioinform.*, 11(3):290–300, 2010.

[4] D. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[5] M. Hao, U. Dayal, D. Keim, and T. Schreck. Multi-resolution techniques for visual exploration of large time-series data. In *Proc. of EuroVis'07*, pages 27–34, 2007.

[6] M. C. Hao, H. Janetzko, S. Mittelstädt, W. Hill, U. Dayal, D. A. Keim, M. Marwah, and R. K. Sharma. A visual analytics approach for peak-preserving prediction of large seasonal time series. *Computer Graphics Forum*, 30(3):691–700, 2011.

[7] C. Holzhüter, A. Lex, D. Schmalstieg, H. Schulz, H. Schumann, and M. Streit. Visualizing uncertainty in biological expression data. In *Proc. of VDA'12*, pages 82940O–82940O–11, 2012.

[8] P. Isenberg, A. Bezerianos, P. Dragicevic, and J.-D. Fekete. A study on dual-scale data charts. *IEEE TVCG*, 17(12):2469–2478, 2011.

[9] W. Javed and N. Elmqvist. Stack zooming for multi-focus interaction in time-series data visualization. In *Proc. of PacificVis'10*, pages 33–40, 2010.

[10] W. Javed and N. Elmqvist. Exploring the design space of composite visualization. In *Proc. of PacificVis'12*, pages 1–8, 2012.

[11] R. Kincaid. SignalLens: Focus+context applied to electronic time series. *IEEE TVCG*, 16(6):900–907, 2010.

[12] G. Li, Y. Wang, L. Zhang, and X. Zhu. Similarity measure for time series based on piecewise linear approximation. In *Proc. of WSCP'09*, 2009.

[13] C. Maus, S. Rybacki, and A. M. Uhrmacher. Rule-based multi-level modeling of cell biological systems. *BMC Sys. Biol.*, 5(166), 2011.

[14] T. Mazza, G. Iaccarino, and C. Priami. Snazer: The simulations and networks analyzer. *BMC Sys. Biol.*, 4(1), 2010.

[15] M. Novotny and H. Hauser. Outlier-preserving focus+context visualization in parallel coordinates. *IEEE TVCG*, 12(5):893–900, 2006.

[16] P. Patel, E. Keogh, J. Lin, and S. Lonardi. Mining motifs in massive time series databases. In *Proc. of ICDM'02*, pages 370–377, 2002.

[17] I. Surovtsova, N. Simus, K. Hubner, S. Sahle, and U. Kummer. Simplification of biochemical models: A general approach based on the analysis of the impact of individual species and reactions on the systems dynamics. *BMC Sys. Biol.*, 6(1), 2012.

[18] K. Takahashi, S. N. V. Arjunan, and M. Tomita. Space in systems biology of signaling pathways – towards intracellular molecular crowding in silico. *FEBS Letters*, 579(8):1783–1788, 2005.

[19] J. J. Tyson. Modeling the cell division cycle: cdc2 and cyclin interactions. *PNAS*, 88(16):7328–7332, 1991.

[20] V. Volterra. Variations and fluctuations of the number of individuals in animal species living together. *J. Cons. Int. Explor. Mer.*, 3(1):3–51, 1928.

[21] D. J. Wilkinson. *Stochastic Modelling for Systems Biology*. Taylor & Francis, 2006.

[22] J. Woodring and H.-W. Shen. Multiscale time activity data exploration via temporal clustering visualization spreadsheet. *IEEE TVCG*, 15(1):123–137, 2009.

[23] J. Zhao, F. Chevalier, and R. Balakrishnan. KronoMiner: Using multi-foci navigation for the visual exploration of time-series data. In *Proc. of CHI'11*, pages 1737–1746, 2011.

[24] J. Zhao, F. Chevalier, E. Pietriga, and R. Balakrishnan. Exploratory analysis of time-series with ChronoLenses. *IEEE TVCG*, 17(12):2422–2431, 2011.

[25] H. Ziegler, M. Jenny, T. Gruse, and D. Keim. Visual market sector analysis for financial time series data. In *Proc. of VAST'09*, pages 83–90, 2010.