Visual analytics of the insulin signalling pathway using phosphorylation time profiles

David Ma[*,†], Sandeep Kaur[*,†], Christian Stolte[**], Michael Bain[*,†] and Seán I. O'Donoghue[**,*]

[*] *Garvan Institute of Medical Research, Sydney, Australia*
[†] *University of NSW, Sydney, Australia*
[**] *CSIRO Computational Informatics, Sydney Australia*

Abstract. Visual analysis of time-course data on protein phosphorylation presents a particular challenge: bioinformatics tools currently available for visualising Omics data in time series have been developed primarily to study gene expression, and cannot easily be adopted to phosphorylation data, where a single protein typically has multiple phosphosites. In this study, we worked with an experimental research group that are applying very recent methods in high-throughput experimental proteomics to study the time course of protein phosphorylation events in human cells in vitro following stimulation by insulin, as part of a broader project on diabetes and obesity. We applied several existing visual analytics approaches with the goal of organizing the data to gain new insight. We developed a novel layout strategy (called 'Minardo'), loosely based on cell topology, but ordered by time and causality. This gives life scientists a familiar and helpful frame of reference for organizing and interpreting these data. The layout proved to be useful, leading to new insight into the insulin response pathway. We are working on generalizing the Minardo layout to accommodate similar datasets related to other signaling pathways, which should be straightforward.

Keywords: Visual analytics, Signalling pathways, Proteomics, Temporal data
PACS: 87.18.Vf

INTRODUCTION

Modern experiments in molecular systems biology are yielding a wealth of detailed data, allowing the behavior of typically tens of thousands of molecular components to be tracked simultaneously. Typical experiments of this kind allow a detailed view of the molecular changes in RNA, protein abundance, or protein posttranslational modifications in response to external stimuli, such as the presence of a hormone. In principle, these data should allow experimentalists to gain fundamental new insights into underlying biological processes - in practise, however, all too often the volume and complexity of the data can be overwhelming, and can obscure discovery.

Since widespread adoption of these high-throughput techniques over the past 15 years, an array of bioinformatics tools have been developed to assist experimentalists in dissecting and understanding such multivariate datasets [1]. While many different clustering, machine learning and other analytical approaches have been developed, the most common approach used to visualize the data and resulting analysis results is a pathway or network representation, usually augmented with multivariate profile plots or heat maps [1]. In these visualizations, a wide range of additional approaches have been developed is to overload the network with data from time-series (or from sets of

different experimental conditions) - typically by using node or edge coloring [1], small multiple views [2], layers [3], or, specifically for time profile data, a range of alternative representations and layouts [4]. In spite of all these innovations, the overall challenges facing experimentalists are largely unsolved, and further innovations are called for [5].

The data sets emerging from the relatively new field of high-throughput proteomics have some particular challenges [6], for example in studying post-translational modifications, each protein can be modified in multiple ways (e.g., phosphorylation, acetylation) and at multiple sites, giving some unique challenges for visual representation. Increasingly often, proteomics experiments are being used to produce time series data, e.g., tracking response after a particular stimulus. However, most bioinformatics tools currently available for visualising Omics data in time series have been developed for gene expression from microarrays, and cannot easily be adopted to phosphorylation time-series dataset, where a single protein can have multiple time-series [7, 8, 9, 10]. A more general and difficult problem is the combined visualization and analysis of phosphorylation data with different kinds of Omics data, especially with protein and transcript abundance, and with different data from experimental conditions and treatments [11].

One strategy for addressing these challenges is the application of principles of visual analytics, which takes advantage of the remarkable visual processing capabilities within the human brain for analyzing patterns and images. Applying such techniques in combination with interactive data visualization and analysis is a promising approach to improving the management and interpretation of the large and complex datasets in the life science [5].

In this study, we worked with an experimental research group that are applying very recent methods in high-throughput experimental proteomics to study the time course of protein phosphorylation events in human cells in vitro following stimulation by insulin [12], as part of a broader project on diabetes and obesity. They had already applied a wide range of existing analysis and visualization tools to these data - however only relatively few tools have been designed with time-course phospho data in mind. Their key unmet requirement was for a system that would enable visual exploration of networks representing insulin response, interactively overlaid with firstly phospho time course data, but later also data on RNA and protein abundance.

We applied several existing visual analytics approaches with the goal of re-organizing the data to gain new insight. We then discussed the merits and weaknesses of the existing approaches with our experimental collaborators, and used this feedback together with visual analytics principles to develop an improved layout strategy, customized for their data. We call our new layout strategy ‘Minardo’, a play on words, as it was partly inspired by the well-known information graphic published by Minard in 1869, showing Napoleon’s disastrous Russian campaign of 1812 - regarded as an exemplar by many data visualization specialists [13]. The Minardo layout revealed several inconsistencies with the published interpretation of this dataset, and suggested several new insights into the timing and order of events underlying the insulin response pathway. While the current layout was constructed specially for analysing phosphorylation data related to insulin response, aspects of the layout have clear potential to be generalized to help with analysing a broad range of systems biology data.

METHODS

Phosphorylation dataset for insulin response

We worked with members of the James laboratory [14] at the Garvan Institute, who are world leading researchers in applying experimental systems biology to study diabetes and obesity. We focused our work on a recently published dataset from this laboratory showing temporal phosphorylation events in proteins in cells that were initially starved, then stimulated with insulin and glucose [12]. These data were obtain by stimulating in vitro mouse cells derived from brown adipose (fat) tissue; after different time points, the cells were lysed and analysed to determine the phosphorylation state of all detectable proteins. The data were collected using mass-spectrometry, which reveal the phoso-states for serine (S), threonine (T) and tyrosine residues (Y) [15], resulting in a final set of time profiles for 7,897 phosphosites that were judged to be of good quality - an average of about 6.5 phosphosites per protein [12].

Humphrey et al. then used unsupervised fuzzy c-means clustering to organize the time profiles for each phosphosite into groups. They also conducting an extensive literature survey to identify the kinases responsible for a subset consisting of 104 of the phospho-sites judged to be most significant, based on prior knowledge of the response pathway. These data are presented in Figure 5 of Humphrey et al. [12] - the data in this figure were used as the starting point for our work, with the goal of re-analysing and organizing these data to provide greater insight into underlying biological processes.

Cerebral layout

We considered the current tools available for combining network visualization and omics data, guided by the review of Gehlenborg et al. [1]; we considered the most promising to be Cerebral [2], a plug-in to the popular Cytoscape framework [16]. Cerebral provides several well-thought-out concepts for visualisation that could be applied to data that combines time-series with a signaling pathway. As with many tools, Cerebral can display time-series data via node coloring overlaid on the network; however, it also provides brushing and linking, enabling both easy visualization of time profile details associated with a single node, as well as highlight of up- and down-stream nodes and edges. In addition, Cerebral uses small multiples, and a layout driven by the sub-cellular topology.

We selected a subset of the dataset corresponding to key proteins and phospho events in the insulin response pathway, then built a network in Cerebral as follows:

1. The network was represented by importing a list of source and target nodes into Cytoscape (version 2.8.2). All of the data were converted into a tabular form and imported into Cytoscape.
2. Since we wished to show the phosphorylation profile for a particular site across the time points, for proteins consisting of multiple sites, sub-nodes representing the specific sites were created stemming from the original node (Figure 1).

3. For each of the nodes, its localisation in the cell, with respect to the events in the signalling pathway was also imported into Cytoscape. This list was compiled using the resources UniProt [17] and the LOCATE subcellular localisation database [18].
4. After all this data was entered into Cytoscape, we created the Cerebral visualisation by selecting 'Create Cerebral view' from the 'plug-ins' menu. Cerebral version 2.8.2 was used for this purpose.
5. The right tabular column titled 'Cerebral' enables the user to set up the visualisation. We defined the 'Expression' and the 'Comparison' colour scale limits to be between 0 and 1, since the data ranges between 0 and 1. We changed the 'Significance cutoff' and the 'Expression cutoff' to be 0.001 and ± 0 respectively.
6. For clarity, Figure 1 shows the visualisation only for a small subset of proteins analyzed.

Heat map of the time series data

We obtained the raw time course data from Humphrey et al. [12], which comes from MaxQuant [19], and typically exists as a large comma-separated text file, which contains the ratios of the absolute values of each time point and a basal level. In this dataset, the basal level is starved cells, and each time point is the amount of time since stimulation with insulin. Thus, we have 9 time points in triplicate, where phosphorylation level at time zero is always set to 1.0 and each value is the ratio of that point's abundance as compared with the basal (for more information, see Methods of Humphrey et al. [12]).

To create a heat map, we averaged the triplicate timepoints and examined each spectra individually, setting the lowest level of activation in the time course to be 0% and the highest level to be 100%. Although log values are often used to allow better statistical comparison between the higher and lower values, we considered them unnecessary for purposes of calculating the heat map, as the data is not naturally exponential, and in this case it would just make the differences between the larger values harder to distinguish.

By drawing the heat map using D3.js we are able to make it interactive. This opens up many possibilities for user interaction, such as: offering different ways of sorting the rows, relabeling the rows, showing/hiding features of the heat map, clicking through to UniProt and various other databases or resources. It also allows us to enable brushing and linking interactions with other graphical elements, such as a pathway diagram. D3 produces scalable vector graphics (SVG), which facilitates interactive zooming, as well as creating detailed print outs in PDF format.

Single time point for each phosphosite

Using the heat map analysis, we devised a method for consistently selecting a representative time point for each phosphosite based on the data. The method for deciding which time point a phosphosite corresponds to is: estimate the time point at which the phosphorylation state of a phosphosite first transitions from below 50% activation to above, or vice versa in the case of a dephosphorylation. We took this timepoint to be

the timepoint of ‘first activation’, in a similar manner to which the Michaelis constant reduces a temporal event into a number.

The time point is calculated as the timepoint which is closest to 50% out of the two timepoints next to which the change occurred. In the case of a tie, the lower % timepoint is used, the reason for this is because the arrow is more visible on lighter colours than darker colours.

Minardo layout

In the Minardo layout, a number of design principles are employed to improve upon concepts in the Cerebral plug-in. The various visual channels available, such as position, hue, connections, size, shape etc. have been carefully chosen to effectively convey information with minimal cognitive load [20, 21].

Firstly, we were motivated to use position to convey the most meaningful information, as it is generally considered the most valuable visual channel. In Minardo we have used the X and Y axis to show time, causality and subcellular topology. This is achieved by dividing the diagram into consecutive sections which are representative of the different time points. We began in Adobe Illustrated, creating a schematic cell, then mapping time in an arc around the cell, divided into the time intervals used to derive the experimental data (Figure 2). With a single time point each, each phosphosite can now be placed unambiguously in one time interval. We also arranged the cell topology such that the regions for each time interval contain extracellular space, cytoplasm, and nuclear space, allowing for positing proteins based on their subcellular location. The concept could be extended to include other more obscure locations, perhaps combined with labels, tooltips and hyperlinks.

Rather than lay out the consecutive time point areas in one direction (say, along the X or Y axes), we have taken inspiration from Charles Joseph Minard’s classic flow map of Napoleon’s March and curled the flow of time around the nucleus, creating a nicer aspect ratio, and allowing connections to be made from later spectra to earlier spectra without increasing the risk of connection overlaps.

Arrows are used to indicate kinases and their target phosphosites. In the current dataset there around 100 such connections. To overcome the typical ‘hairball’ problem that occurs with networks of around this size and larger, we reduced clutter by using tracks to represent ‘promiscuous’ proteins or complexes, i.e., those involved in multiple phosphorylation events in multiple time points. This is similar to the concept of hubs, or high-degree nodes of a network, but modified to account for the time-course dataset.

To reduce overloading the hue channel, hue was used consistently in both the heat map and the network, with red, green, and blue were used to represent Serine, Threonine and Tyrosine residues, respectively. Yellow is used to indicate information that is currently highlighted, but sometimes we are interested in highlighting more than one type of information at the same time. To prevent overload, the default highlight is Yellow only, and it shows the relevant kinases and phosphorylation events on a track or the phosphates currently being brushed over. ‘Show Targets’ is then given a toggle button, which turns Teal when switched on, this lets the user know that targets are now being shown with a

Teal highlight.

The layout was saved in SVG, and imported into an HTML page, together the with heat map. JavaScript was used to create interactive brushing and linking between the two representations.

Using 3D structure information

In our initial Minardo layout (Figure 2), phosphosites are only represented only using protein name and residue numbering. This helps create a clean, sparse layout, but in some cases it can be useful to see more information about a protein. For example, seeing an image corresponding to a protein's 3D structure can often give biologists insight into its function [22]. To test the utility of such representations, we calculated images for all proteins in the dataset, using residue coloring to indicate the location of each phosphosite, then tried adding these images to the network layout. The space filling models were generated using QuteMol [23] from Protein Database (PDB) files, using the PSSH database [24] to find solved structures that are most closely related to our protein of interest. Each 3D model was manually rotated so that the final 2D view shows all phosphosites, if possible.

RESULTS

Cerebral layout

We initially used the Cytoscape plug-in 'Cerebral' to visualize part of the Humphrey et al. dataset (Figure 1). We then collected user feedback on strengths and weakness of the layout from members of the experimental group that had produced the original dataset. Features of the layout judged by the users to be positive included:

- the use of small multiples to show the state of the pathway at different time points;
- brushing and linking between a node in the pathway and its phosphorylation profile;
- brushing a node to show and downstream events in pathway.

On the negative side, we found difficulties in mapping our data onto the Cerebral layout, which uses the Y-axis to show a combination of subcellular compartment and causal flow. This layout concept seems to work well for showing gene expression data, where a signal traverses from the cell exterior into the nucleus, resulting in downstream changes to transcript abundance. However, for the current phospho time series data, where most of the action takes place within the cytoplasm, it was not clear how to arrange the data along the Y-axis. Additionally, we found that the approach of using one node per phosphosite meant that when constructing a pathway of around 100 sites, the view quickly became visually cluttered. We noted a number of technical limitations when using the current Cerebral plug-in:

- Clicking on a node in Cytoscape accesses URLs with more specific information, however Cerebral currently disables this feature.

- There were problems with zooming that make the visualization unusable on small laptops.

A practical limitation we found in using any specific Cytoscape plug-in was the difficulty in sharing the interactive view with our experimental collaborators, since each user needs to download the correct software versions, and needs to learn basic Cytoscape operations before they can reconstruct the same view.

A more central concern with this layout was expressed by the users we interviewed: most of the screen is devoted to displaying a single model which tries to show the overall flow of causal events - however, the underlying experimental data are divided fundamentally into distinct time points. Although the time data is shown in the small multiples, the users reported that it was somewhat difficult to infer functional activity, as it requires going back and forward between the main pathway and the small multiple views. In summary, we realized that the view did not make best use of the key spatial dimensions (X & Y) to show directly the time-series data.

Minardo layout

To address some of the key issues identified above, while preserving the brushing and linking feature, identified as very useful by the users, as well as preserving aspects of the small multiple viewers, we created an interactive HTML file showing the time-series data in a heat map linked with a 'Minardo' network layout (see Figure 2). The network layout shows the temporal and causal order of phosphorylation events (taken from Fig. 5 of Humphrey et al. [12]), with arrows identifying each kinase and its substrate phosphosite. The proteins Akt, Irs1, and AS160, p70S6K, Erk1/2 and the complexes Gsk, mTORC1, and mTORC2 play roles across multiple times, so are indicated with white tracks running parallel to the membranes. The HTML file containing the complete figure is included in the Supplementary Information; it allows searching for proteins by name (works best in the Safari browser), as well as brushing and linking between the heat map and the network - i.e., hovering over a protein name causes automated highlighting of all occurrences of the name. Also, hovering over protein names in the heat map opens a small overlay which shows the complete protein name, and provides a link to the corresponding UniProt entry, providing comprehensive information about the protein and its function. We also provided buttons to allow the user to sort the heat map according to different criteria.

By making the HTML version available within our organization's intranet, we were easily able to disseminate it to our users and get their feedback. Overall, the user feedback was very positive. The brushing and linking between the heat map and network was noted as being very helpful for interpreting the data in detail. However, the most positive feedback was that the new layout helped them gain important new insight from their dataset into important aspects of the underlying biological processes. In particular, the Minardo layout made the following points clear:

1. Both Plin1 (T198) and MEKK3 (S166) are very rapidly phosphorylated, with over 50% of sites switched before the 15 s time point. In Fig. 5 of Humphrey et al.,

these phosphorylations are indicated as caused by the kinases PKA and SGK1, respectively, with SGK1 in term phosphorylated by the kinase PDK1. However, the current dataset does not indicate that neither PKA, SGK1, or PDK1 are activated that early. The new layout suggests an alternative hypothesis, that, like Irs1 (Y465 & Y1179), Plin1 (T198) and MEKK3 (S166) are phosphorylated early, either by Igfr1, or perhaps by another kinase not identified.

2. Interestingly, all of the four most rapidly phosphorylated sites MEKK2 (S166), Plin1 (T198), and Irs1 (Y465 & Y1179), which change from 0 to 100% in 15 s, are also all rapidly dephosphorylated, falling significantly by 30 s. The new layout suggests not only a common phosphorylation switch for these four sites, but also a common dephosphorylation switch. If validated, these observations could be a significant step forward in our understanding of insulin response, and may indicate that MEKK2 and Plin1 play key roles not yet understood.
3. The dissociation of Akt from the plasma membrane is believed to require phosphorylation of both T308 and S473. In Figures 5 and 6 from Humphrey et al., dissociated Akt is shown to phosphorylate substrates Gsk3b (S9), Bad (S99), AS160 (T652, S318, & S588), El24 (S47), FOXO1A (S316), and Tsc2 (T1462). The new Minardo layout however suggest that this interpretation is unlikely, since these events occur prior to 1 minute, which is when 50% of Akt (S473) sites are phosphorylated. Instead, the new layout implies that either these substrates are phosphorylated by Akt while it is still bound to the plasma membrane, or by by other kinases - perhaps either Igf1r, in direct response to insulin binding, or by the Gsk or mTORC2 complexes, which both get phosphorylated by 15 s.
4. Humphrey et al. indicated that phosphorylation of PKA (T198) plays to a key role in initiating lipid metabolism by initiating subsequent phosphorylation of Plin1 (S497), HSL (S855 & S951), and TAK1 (S439). With the new layout, this seems an unlikely explanation for the rapid phosphorylation of Plin1 (T198). Overall, PKA (T198) shows a general increases in phosphorylation up until a maximum at 2 mins; however, in a marked contrast to the previous indication, over this time both HSL (S855) and TAK1 (S439) become progressively dephosphorylated. Also, at around the same time that PKA is completely dephosphorylated, HSL (S951) becomes rapidly phosphorylated (5 min).

These above observations were judged by our users to give quite significant new information on insulin response, and we are now planning a joint publication with them in a biological journal.

While the users found the current layout useful, they also requested features that it does not yet support. First and foremost was the ability to automatically modify which phosphosites are used to construct the layout - the current dataset shows only 104 of the 7,897 phosphosites judged to be of good quality. Secondly, the users would like to be able to interactively edit the graph to change the kinase and target assignments. Thirdly, the users were interested in combining other data with this layout, including protein abundance and multiple experimental conditions (such as the presence of various chemical inhibitors). Finally, users were interested in adding a more fully capable search feature that can match different synonyms for the same protein, since many proteins are known by multiple names (e.g. As160, Kiaa0603, Tbc1d4 all refer to the same protein).

Using 3D structure information

Where available, we added images of 3D structures for proteins in one region of the layout (Figure 3), marking the phospho residues using the same colouring scheme as in the heat map (red = serine, blue = tyrosine, and green = threonine). Phosphosites are usually solvent accessible, and visualizing them can give insight that helps to validate them as substrates [25]. The goal was to provide additional information that can assist the user in a more detailed interpretation of the experimental dataset. Feedback from our users suggested that, where available, such information is indeed useful in this context; they also commented favorably on other aspects of the visualization that added biological context, such as the stylized representations of lipid bilayers and subcellular compartments. However, a key limitation was that many of the phosphosites occur in regions that have no detectable sequence similarity to proteins with known structure - for the complete dataset, this occured in about half of the 104 sites. This is consistent with the observation that many phosphosites occur in intrinsically disordered region, which are hard to determine structurally [26].

DISCUSSION

The new layout strategy and its implement to the insulin response pathway has provided a framework for re-interpreting the current dataset, as well as combining it with other related data, such as 3D structural information. In this case, application of visual analytic principles, together with close feedback from the experimentalists who generated the data, has lead to several new insights into the detailed mechanisms behind insulin response. This provides a demonstration of how the analysis and interpretation of such complex data is sometimes a major bottleneck in on molecular systems biology, and how visual analytics may help.

A key factor behind the Minardo layout is to map both time and causality onto the graph - this allows for spatial reasoning about separate stages of the experiment, effectively subdividing the full dataset into more manageable chunks. A second key element is the identification of promiscuous proteins and complexes - i.e., those involved in phospho events at multiple time intervals - and representing them as parallel tracks that span multiple time intervals. This process is somewhat akin to parallel coordinate visualization [27], where single points become lines. The resulting tracks are then represented in a form that is visually very distinct from the edges that represent phopho events; this results in a graph with no apparent edge crossing, a very beneficial outcome for helping users understand and interpret information in the graph [28].

While the existing method has helped with the current dataset, some clear shortcomes were identified that we plan to address. As indicated above, the key next steps will be to automate the layout to facilitate easy analysis of other phospho time course data. We plan to address the protein synonym issue mentioned by using the Reflect resource [29], a widely-used system that maintains a community-edited dictionary of protein synonyms.

One limitation of the current design is that the order of events shown within one time interval imply an ordering in time, which is incorrect. In future versions, we plan to investigate strategies to address this limitation, e.g., basing order on more precise

estimates derived from the time series.

White the Minardo layout appears to be an improvement on existing solution to proteomics time-course data, it currently does not address the harder and more general problems of how to show multiple conditions, or protein and transcript abundance. As a further extension of this work, we plan to test if our approach can accommodate some of the existing approaches to visualizing these additional data, such as node colouring or decoration [1].

Also as part of future work we plan to investigate combining our visualization strategy with more advanced qualitative modelling methods, adding constraints determined from prior biological knowledge and applying machine learning to reverse engineering the functional events in the signalling pathway. The problem of deducing an underlying phosphorylation network based on datasets such as used in this work can be described using intrinsically qualitative formulations. Such formulations allowing the use of well-established methods to determine properties of states and trajectories, such as the identification of cycles, or unreachable states [30]; in addition, they enable the use of network inference algorithms to 'reverse engineer' hypothesised systems models from data [31]. We are interested in applying methods to identification from a very large space of 'phospho-forms' those most likely to be present in the data [32], by applying existing knowledge-based techniques that can make use of such constraints [33].

Conclusions

The Minardo layout provides a novel combination of principles from visual analytics in a customized layout loosely based on cell topology, but strictly ordered first by time, then by causality. This gives life scientists a familiar and helpful frame of reference for organizing and interpreting proteomics time-course data. The layout has proven to be useful, leading to new insight into the insulin response pathway. We are working on generalizing the Minardo layout to accommodate similar datasets related to other signaling pathways, which should be straightforward.

ACKNOWLEDGMENTS

We gratefully acknowledge helpful conversations with our colleagues Prof. David James, Dr Sean Humphrey, Annabel Minard, Beverley Murrow.

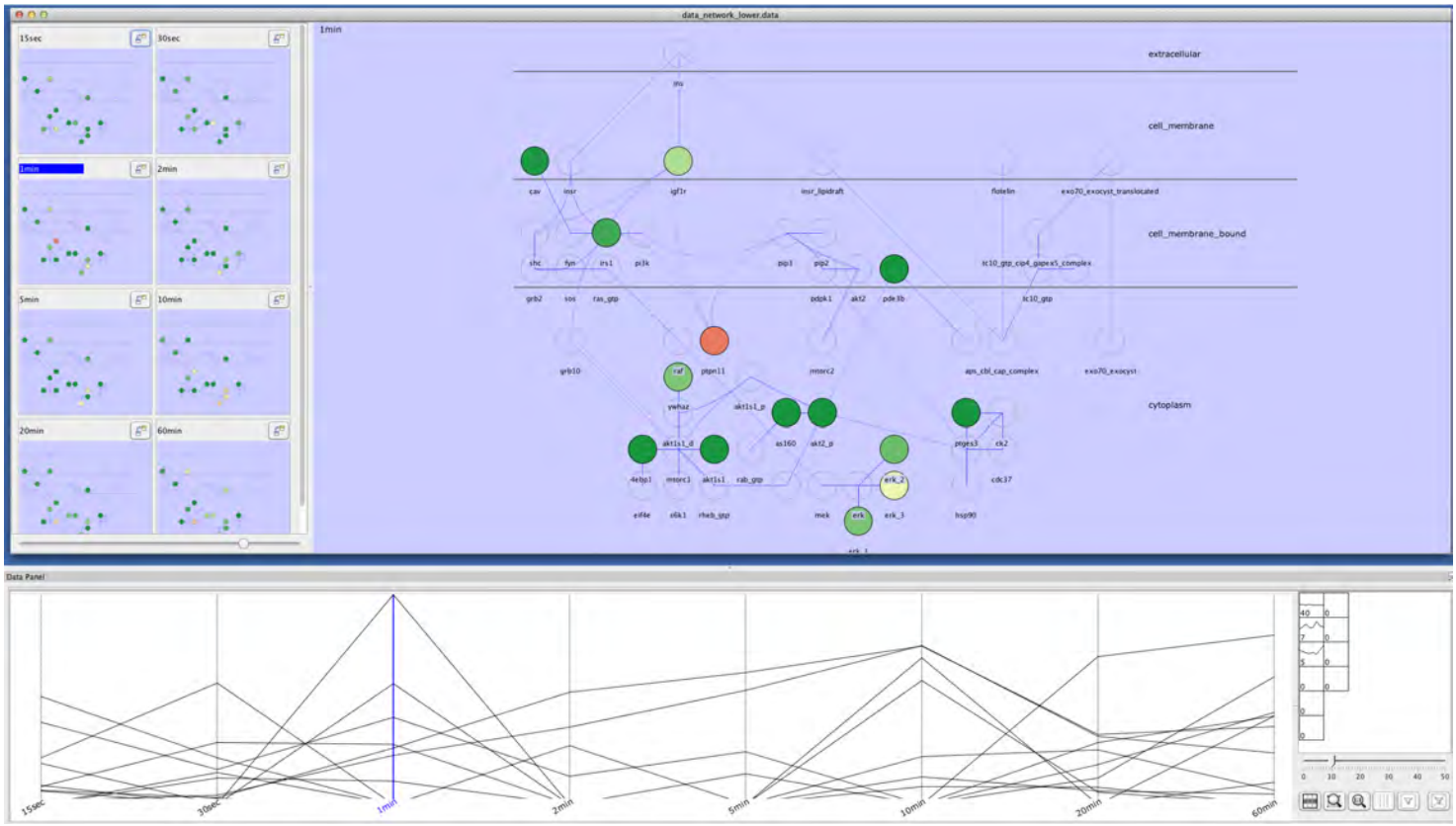

FIGURE 1. Cerebral layout of insulin response dataset. Shown is a subset of the insulin response data visualized using the Cytoscape plug-in Cerebral [2]. When a protein has multiple phosphosites, each are shown using a separate sub-node stemming from the original node for that protein. The state of the network at eight different time point is indicated with small multiples (top left), where node colors indicate high (green) or low (orange/brown) levels of phosphorylation. The user can examine one time point in detail by clicking on its small multiple representation, causing it to become loaded in the main view. Nodes in the main view are connected to the profile plot (bottom) using brushing and linking, making it easy to see the exact phosphorylation profile for any given phosphosite. Nodes in the main view are arrange to show a flow of causality from top to bottom, and simultaneously a transition between different subcellular compartments.

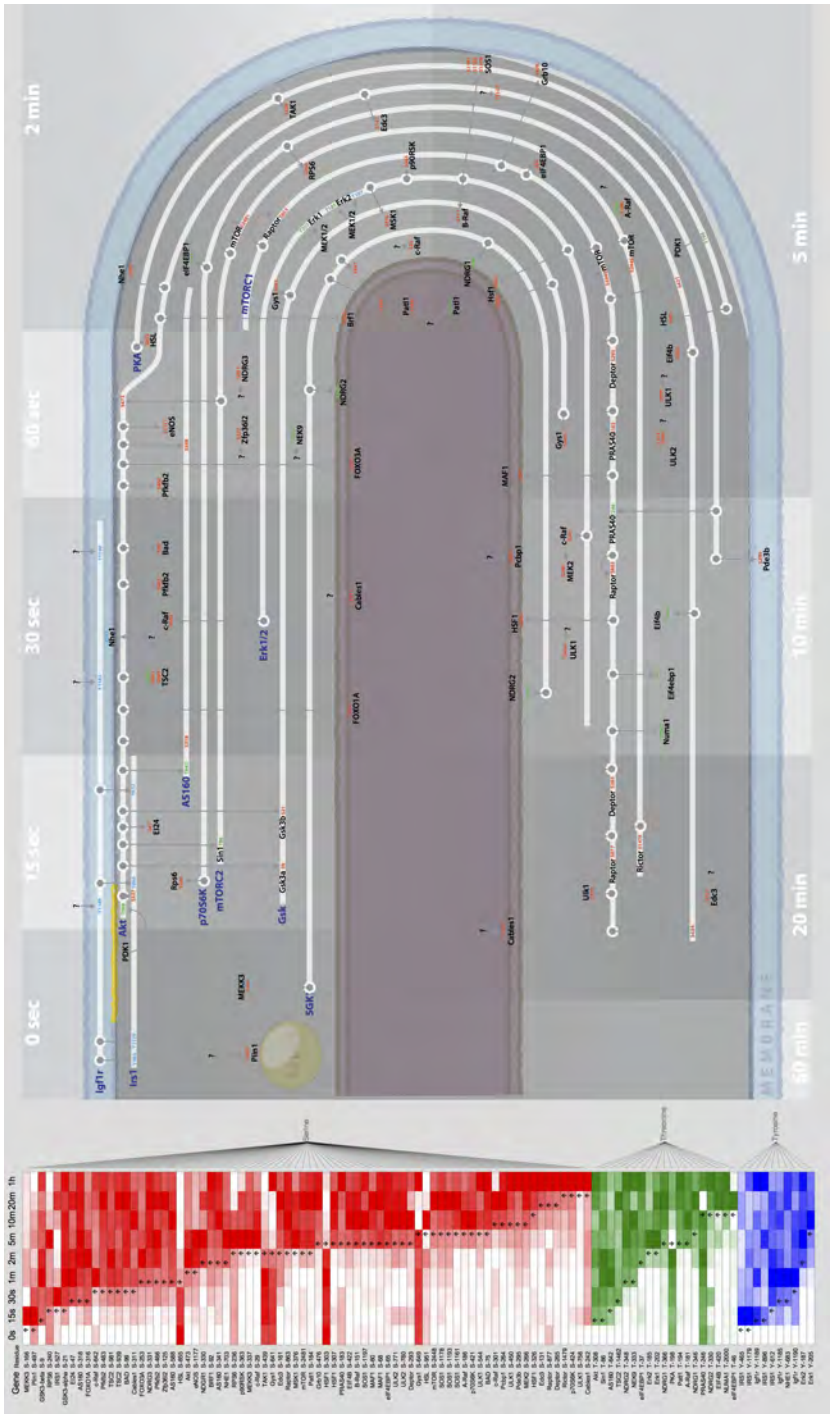

FIGURE 2. Heat map and Minardo layout of insulin response dataset. The heat map (left) shows time profiles for 104 phosphosites (red for Serine, green for Threonine, blue for Tyrosine), ordered by first half-maximal change. The Minardo layout (right) shows the temporal and causal order of phosphorylation events; each arrow identifies a kinase and its substrate phosphosite. The proteins Akt, Irs1, and AS160, p70S6K, Erk1/2 and the complexes Gsk, mTORC1, and mTORC2 play roles across multiple times, so are indicated with white tracks running parallel to the membranes. Akt and Irs1 are both initially bound to the inner plasma membrane, but is shown to gradually disassociate from the membrane after around 2 minutes - this process is well established [34, 12], although not the precise timing. Phosphorylation of the protein SGK was not observed in this dataset but has been added based on prior published work [12]. An HTML version of the complete figure is included in the Supplementary Information; it allows search by name, as well as brushing and linking between the heat map and the network, which helps users make detailed interpretations from the dataset.

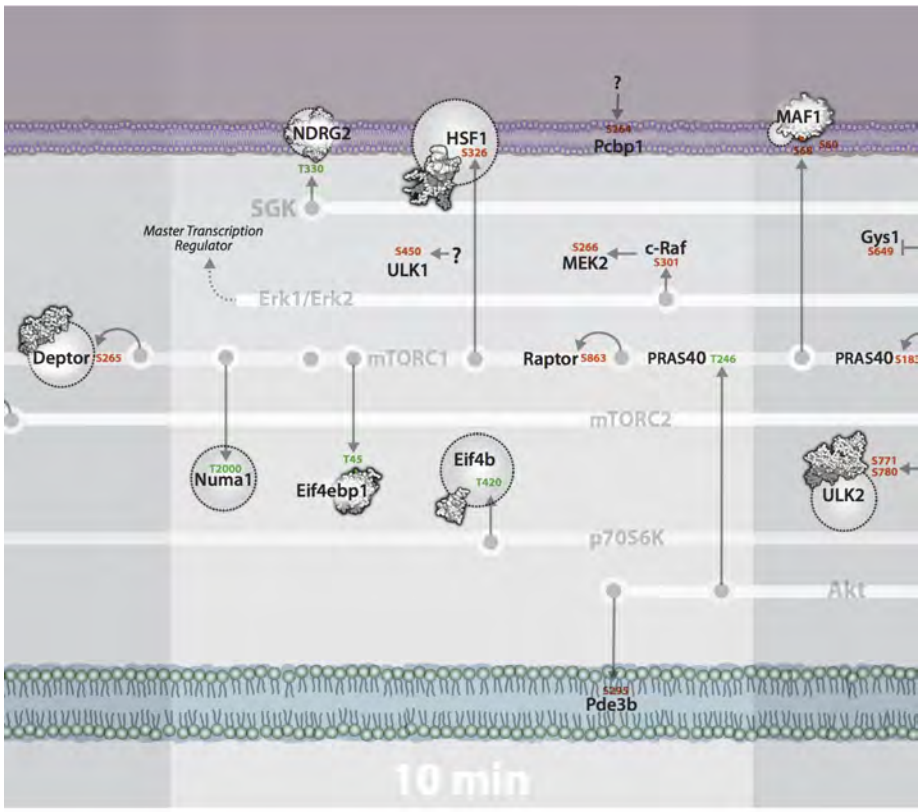

FIGURE 3. Detail from Minardo layout. Detail of the layout in Figure 2, augmented to show, where available, space-filling models of each protein, indicating the location of each phosphosite. Phospho residues are indicated using the same colouring scheme as in the heat map in Figure 2 (red = serine, blue = tyrosine, and green = threonine).

REFERENCES

1. N. Gehlenborg, S. O'Donoghue, N. Baliga, A. Goesmann, M. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, D. Tenenbaum, and A.-C. Gavin, *Nature Methods* **7**, S56–S68 (2010).
2. A. Barsky, J. Gardy, R. Hancock, and T. Munzner, *Bioinformatics* **28**, 1040 – 1042 (2007).
3. G. A. Pavopoulos, S. I. O'Donoghue, V. P. Satagopam, T. G. Soldatos, E. Pafilis, and R. Schneider, *BMC Systems Biology* **2**, 104 (2008).
4. M. Secrier, and R. Schneider, *Brief Bioinform.* (2013).
5. S. I. O'Donoghue, A.-C. Gavin, N. Gehlenborg, D. S. Goodsell, J.-K. Hériché, C. B. Nielsen, C. North, A. J. Olson, J. B. Procter, D. W. Shattuck, T. Walter, and B. Wong, *Nature Methods* **7**, S2–S4 (2010).
6. M. Bessarabova, A. Ishkin, L. JeBailey, T. Nikolskaya, and Y. Nikolsky, *BMC Bioinformatics* **13**, 104 (2012).
7. D. Emig, N. Salomonis, J. Baumbach, T. Lengauer, B. R. Conklin, and M. Albrecht, *Nucleic Acids Res.* **38**, W755 – W762 (2010).
8. T. Xia, J. V. Hemert, and J. A. Dickerson, *Bioinformatics* (2010).
9. T. Wittkop, D. Emig, S. Lange, S. Rahmann, M. Albrecht, J. H. Morris, S. Böcker, J. Stoye, and J. Baumbach, *Nature Methods* **7**, 419 – 420 (2010).
10. E. Bonnet, L. Calzone, D. Rovera, G. Stoll, E. Barillot, and A. Zinovyev, *BMC Systems Biology* **7** (2013).
11. R. Saito, M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, S. Lotia, A. R. Pico, G. D. Bader, and T. Ideker, *Nature Methods* **9**, 1069 – 1076 (2012).
12. S. Humphrey, G. Yang, P. Yang, D. Fazakerley, J. Stockli, J. Yang, and D. James, *Cell metabolism* **17**, 1–12 (2013).
13. M. Bessarabova, A. Ishkin, L. JeBailey, T. Nikolskaya, and Y. Nikolsky, *The Visual Display of Quantitative Information*, Graphics Press, 2001, 2nd edn.
14. James lab (2013), URL `http://www.jameslab.com.au`.
15. P. Cohen, *Nature Cell Biology* **4**, E127–130 (2002).
16. P. Shannon, A. Markiel, O. Ozier, N. Baliga, J. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, *Genome research* **13**, 2498 – 2504 (2003).
17. M. Margrane, and the UniProt consortium, *Database* (2011).
18. J. Sprenger, J. L. Fink, S. Karunaratne, K. Hanson, N. Hamilton, and R. Teasdale, *Nuclic acids research* **36**, D230 – D233 (2008).
19. J. Cox, and M. Mann, *Nat. Biotechnol.* **26**, 1367 — 1372 (2008).
20. J. Sweller, *Learning and Instruction* **4**, 295–312 (1994).
21. P. Chandler, and J. Sweller, *Cognition and Instruction* **8**, 293–332 (1991).
22. S. I. O'Donoghue, D. S. Goodsell, A. S. Frangakis, F. Jossinet, R. Laskowski, M. Nilges, H. R. Saibil, A. Schafferhans, R. Wade, E. Westhof, and A. J. Olson, *Nature Methods* **7**, S42–S55 (2010).
23. M. Tarini, P. Cignoni, and C. Montani, *IEEE Transactions on Visualization and Computer Graphics* **12**, S42–S55 (2006).
24. I. Schafferhans, J. E. W. Meyer, and S. I. O'Donoghue, *Nucleic Acids Res.* **31**, 494–498 (2003).
25. A. Zanzoni, D. Carbajo, F. Diella, P. F. Gheradini, A. Tramontano, M. Helmer-Citterich, and A. Via, *Nucleic Acids Res.* **39**, D268–D271 (2011).
26. T. Chouard, *Nature* **471**, 151–153 (2011).
27. A. Inselberg, *Visual Computer.* **1**, 69—91 (1985).
28. C. Ware, H. Purchase, L. Colpoys, and M. McGill, *Information Visualization* **1**, 103–110 (2002).
29. E. Pafilis, S. I. O'Donoghue, L. J. Jensen, M. Kuhn, N. P. Brown, and R. Schneider, *Nature Biotechnology* **27**, 308–310 (2009).
30. R. David, and H. Alla, *Discrete, Continuous, and Hybrid Petri Nets*, Springer, Berlin, 2010, Second edn.
31. R. Samaga, J. Saez-Rodriguez, L. Alexopoulos, P. Sorger, and S. Klamt, *PLoS Comput Biol* **5**, e1000438 (2009).
32. S. Prabakaran, R. Everley, J. Landrieu, I. Wieruszeski, G. Lippens, and J. Steen, H. Gunawardena, *Mol Syst Biol* **7** (2011).
33. A. Srinivasan, and M. Bain, "Knowledge-Guided Identification of Petri Net Models of Large Biological Systems," in *Proc. 21st Intl. Conference on Inductive Logic Programming (ILP 2011; Revised*

Selected Papers), edited by S. Muggleton, A. Tamaddoni-Nezhad, and F. Lisi, Springer, Berlin, 2012, vol. 7207 of *Lecture Notes in Computer Science*, pp. 317–331.
34. R. A. Heller-Harrison, M. Morin, and M. P. Czech, *J Biol Chem.* **270**, 24442–24450 (1995).