Advances in Experimental Medicine and Biology 823

Changming Sun
Tomasz Bednarz
Tuan D. Pham
Pascal Vallotton
Dadong Wang *Editors*

Signal and Image Analysis for Biomedical and Life Sciences

Springer

Advances in Experimental Medicine and Biology

Volume 823

More information about this series at http://www.springer.com/series/5584

Changming Sun • Tomasz Bednarz • Tuan D. Pham
Pascal Vallotton • Dadong Wang
Editors

Signal and Image Analysis for Biomedical and Life Sciences

Springer

Editors
Changming Sun
Tomasz Bednarz
Pascal Vallotton
Dadong Wang
Digital Productivity Flagship, CSIRO
Sydney, NSW, Australia

Tuan D. Pham
The University of Aizu
Fukushima, Japan

ISSN 0065-2598 ISSN 2214-8019 (electronic)
ISBN 978-3-319-10983-1 ISBN 978-3-319-10984-8 (eBook)
DOI 10.1007/978-3-319-10984-8
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014955163

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

With an emphasis on applications of computational models for solving modern challenging problems in biomedical and life sciences, this book aims to bring collections of articles from biologists, medical/biomedical and health science researchers together with computational scientists to focus on problems at the frontier of biomedical and life sciences. The goals of this book are to build interactions of scientists across several disciplines and to help industrial users apply advanced computational techniques for solving practical biomedical and life science problems.

This book is for users in the fields of biomedical and life sciences who wish to keep abreast with the latest techniques in signal and image analysis. The book presents a detailed description to each of the applications. It can be used by those both at graduate and specialist levels.

We have included 14 chapters in this book. Some of the chapters are extensively revised versions of papers that were presented at the International Symposium on Computational Models for Life Sciences held on 27–29 November 2013 in Sydney, Australia. There are two main parts in the book: signal and image analysis issues within the subjects of the book.

In the first part of the book, Chap. 1 presents a novel visualisation strategy tailored for proteomics data. A dataset is visualised showing phosphorylation events in response to insulin that leads to new insights into the insulin response pathway. A strategy for web-based presentation of data is also described. Chapter 2 proposes a new approach for the modelling of testosterone regulation to identify all model parameters from the hormone concentrations of testosterone and luteinizing hormone. Simulation results are described to reveal behaviour similar to clinical data. Chapter 3 proposes two distinct hybrid algorithms that combine efficient sequential change-point detection procedures with the Cross-Entropy method. Results show effectiveness of the described method. In Chap. 4, two methods for distinguishing between healthy controls and patients diagnosed with Parkinson's disease by means of recorded smooth pursuit eye movements are presented and evaluated. The results are indicative of the potential of the presented methods as diagnosing or staging

tools for Parkinson's disease. Chapter 5 presents an approach for the identification of the Reichardt elementary motion detector model. A pool of spatially distributed elementary motion detectors is considered, and a way of designing the visual stimuli for a certain order of spatial resolution is suggested. Chapter 6 discusses on the complexity ensemble measures for gait time series analysis that could have a significantly wider application scope ranging from diagnostics and early detection of physiological regime change to gait-based biometrics. Chapter 7 presents the development of a motion capturing and load analyzing system for caregivers aiding a patient to sit up in bed. The difference between the performances of the two types of caregivers were found: the professional adopted a posture that was safe and did not stress the lumbar vertebrae, whereas the layperson tended to adopt an unsafe posture. Chapter 8 proposes an unsupervised multi-scale K-means algorithm to distinguish epileptic EEG signals and identify epileptic zones. The experimental results demonstrate that identifying seizure with multi-scale K-means algorithm and delay permutation entropy achieves higher accuracy than that of K-means and support vector machine. Chapter 9 presents a method for the tracking of EEG activity using motion estimation in brain topomaps to understand the mechanism of brain wiring. Authors demonstrate that it is possible to track the path of a signal across various lobes.

In the second part of the book, Chap. 10 presents an approach to processing ultra high-resolution, large-size biomedical imaging data for the purposes of detecting and quantifying vasculature and microvasculature. The results on cerebral and liver vasculatures of a mouse captured at the Shanghai Synchrotron Radiation Facility are presented. Chapter 11 describes a novel way of carrying out image analysis, reconstruction and processing tasks using cloud based service provided on the Australian National eResearch Collaboration Tools and Resources infrastructure. The toolbox is available on the web. Chapter 12 presents an investigation into how Massey University's Pollen Classifynder can accelerate the understanding of pollen and its role in nature. Chapter 13 presents a digital image processing and analysis approach for activated sludge wastewater treatment. Chapter 14 presents a complete system for 3D reconstruction of roots grown in a transparent gel medium or washed and suspended in water.

We thank all the authors for their contributions to this edited book. We also thank Dan Hills and Susan McMaster from CSIRO Contracts and Legal for their help with the Publishing Agreement between Springer and CSIRO. We are grateful to Dr. Thijs van Vlijmen, Sara Germans-Huisman, Magesh Kaarthick Sundaramoorthy, and other editors at Springer and S. Madhuriba at SPi Technologies India Private Ltd. for their help and great support from the beginning to the production of this book. Materials from the American Institute of Physics (AIP) Publishing that are used by some authors are acknowledged and credits are given in the respective chapters within this book.

Sydney, Australia — Changming Sun
Tomasz Bednarz
Aizu, Japan — Tuan D. Pham
Sydney, Australia — Pascal Vallotton
Dadong Wang

July 2014

Contents

Part I Signal Analysis

1 Visual Analytics of Signalling Pathways Using Time Profiles 3
David K.G. Ma, Christian Stolte, Sandeep Kaur, Michael Bain, and Seán I. O'Donoghue

2 Modeling of Testosterone Regulation by Pulse-Modulated Feedback 23
Per Mattsson and Alexander Medvedev

3 Hybrid Algorithms for Multiple Change-Point Detection in Biological Sequences 41
Madawa Priyadarshana, Tatiana Polushina, and Georgy Sofronov

4 Stochastic Anomaly Detection in Eye-Tracking Data for Quantification of Motor Symptoms in Parkinson's Disease 63
Daniel Jansson, Alexander Medvedev, Hans Axelson, and Dag Nyholm

5 Identification of the Reichardt Elementary Motion Detector Model 83
Egi Hidayat, Alexander Medvedev, and Karin Nordström

6 Multi-complexity Ensemble Measures for Gait Time Series Analysis: Application to Diagnostics, Monitoring and Biometrics 107
Valeriy Gavrishchaka, Olga Senyukova, and Kristina Davis

7 Development of a Motion Capturing and Load Analyzing System for Caregivers Aiding a Patient to Sit Up in Bed 127
Akemi Nomura, Yasuko Ando, Tomohiro Yano, Yosuke Takami, Shoichiro Ito, Takako Sato, Akinobu Nemoto, and Hiroshi Arisawa

8 Classifying Epileptic EEG Signals with Delay Permutation Entropy and Multi-scale K-Means 143
Guohun Zhu, Yan Li, Peng (Paul) Wen, and Shuaifang Wang

9 Tracking of EEG Activity Using Motion Estimation to Understand Brain Wiring 159
Humaira Nisar, Aamir Saeed Malik, Rafi Ullah, Seong-O Shim, Abdullah Bawakid, Muhammad Burhan Khan, and Ahmad Rauf Subhani

Part II Image Analysis

10 Towards Automated Quantitative Vasculature Understanding via Ultra High-Resolution Imagery 177
Rongxin Li, Dadong Wang, Changming Sun, Ryan Lagerstrom, Hai Tan, You He, and Tiqiao Xiao

11 Cloud Based Toolbox for Image Analysis, Processing and Reconstruction Tasks 191
Tomasz Bednarz, Dadong Wang, Yulia Arzhaeva, Ryan Lagerstrom, Pascal Vallotton, Neil Burdett, Alex Khassapov, Piotr Szul, Shiping Chen, Changming Sun, Luke Domanski, Darren Thompson, Timur Gureyev, and John A. Taylor

12 Pollen Image Classification Using the Classifynder System: Algorithm Comparison and a Case Study on New Zealand Honey 207
Ryan Lagerstrom, Katherine Holt, Yulia Arzhaeva, Leanne Bischof, Simon Haberle, Felicitas Hopf, and David Lovell

13 Digital Image Processing and Analysis for Activated Sludge Wastewater Treatment 227
Muhammad Burhan Khan, Xue Yong Lee, Humaira Nisar, Choon Aun Ng, Kim Ho Yeap, and Aamir Saeed Malik

14 A Complete System for 3D Reconstruction of Roots for Phenotypic Analysis 249
Pankaj Kumar, Jinhai Cai, and Stanley J. Miklavcic

Index 271

Contributors

Yasuko Ando Yokohama City University, Yokohama, Japan

Hiroshi Arisawa Yokohama National University, Yokohama, Japan

Yulia Arzhaeva Digital Productivity Flagship, CSIRO, North Ryde, Sydney, NSW, Australia

Hans Axelson Department of Neuroscience, Neurophysiology, Uppsala University, Uppsala, Sweden

Michael Bain The University of NSW, Sydney, NSW, Australia

Abdullah Bawakid Faculty of Computing and Information Technology, King Abdul Aziz University, Jeddah, Kingdom of Saudi Arabia

Tomasz Bednarz Digital Productivity Flagship, CSIRO, Sydney, NSW, Australia

Leanne Bischof Digital Productivity Flagship, CSIRO, North Ryde, Sydney, NSW, Australia

Neil Burdett Digital Productivity Flagship, CSIRO, Brisbane, QLD, Australia

Jinhai Cai School of Information Technology and Mathematical Sciences, Phenomics and Bioinformatics Research Centre, Australian Centre for Plant Functional Genomics, University of South Australia, Mawson Lakes, SA, Australia

Shiping Chen Digital Productivity Flagship, CSIRO, Sydney, NSW, Australia

Kristina Davis Department of Pathology, University of Michigan, Ann Arbor, MI, USA

Luke Domanski CSIRO IM&T, Sydney, NSW, Australia

Valeriy Gavrishchaka Department of Physics, West Virginia University, Morgantown, WV, USA

Timur Gureyev CSIRO Manufacturing Flagship, Melbourne, VIC, Australia

Simon Haberle School of Culture, History and Language, The Australian National University, Canberra, ACT, Australia

You He Shanghai Synchrotron Radiation Facility (SSRF), Chinese Academy of Sciences, Shanghai Institute of Applied Physics, Shanghai, China

Egi Hidayat Department of Information Technology, Uppsala University, Uppsala, Sweden

Katherine Holt Institute of Natural Resources, Massey University, Palmerston North, New Zealand

Felicitas Hopf School of Culture, History and Language, The Australian National University, Canberra, ACT, Australia

Shoichiro Ito Yokohama National University, Yokohama, Japan

Daniel Jansson Department of Information Technology, Uppsala University, Uppsala, Sweden

Sandeep Kaur Garvan Institute of Medical Research, Sydney, NSW, Australia

The University of NSW, Sydney, NSW, Australia

Muhammad Burhan Khan Faculty of Engineering and Green Technology, Department of Electronic Engineering, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia

Alex Khassapov CSIRO IM&T, Melbourne, VIC, Australia

Pankaj Kumar School of Information Technology and Mathematical Sciences, Phenomics and Bioinformatics Research Centre, Australian Centre for Plant Functional Genomics, University of South Australia, Mawson Lakes, SA, Australia

Ryan Lagerstrom Digital Productivity Flagship, CSIRO, North Ryde, Sydney, NSW, Australia

Xue Yong Lee Faculty of Engineering and Green Technology, Department of Electronic Engineering, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia

Rongxin Li Digital Productivity Flagship, CSIRO, North Ryde, Sydney, NSW, Australia

Yan Li Faculty of Health, Engineering and Sciences, University of Southern Queensland, Toowoomba, QLD, Australia

David Lovell Digital Productivity Flagship, CSIRO, Canberra, Australia

David K.G. Ma Garvan Institute of Medical Research, Sydney, NSW, Australia

The University of NSW, Sydney, NSW, Australia

Aamir Saeed Malik Department of Electrical and Electronic Engineering, Centre for Intelligent Signal and Imaging Research, Universiti Teknologi PETRONAS, Tronoh, Perak, Malaysia

Per Mattsson Department of Information Technology, Uppsala University, Uppsala, Sweden

Alexander Medvedev Department of Information Technology, Uppsala University, Uppsala, Sweden

Stanley J. Miklavcic School of Information Technology and Mathematical Sciences, Phenomics and Bioinformatics Research Centre, Australian Centre for Plant Functional Genomics, University of South Australia, Mawson Lakes, SA, Australia

Akinobu Nemoto Yokohama City University, Yokohama, Japan

Choon Aun Ng Faculty of Engineering and Green Technology, Department of Environmental Engineering, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia

Humaira Nisar Faculty of Engineering and Green Technology, Department of Electronic Engineering, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia

Akemi Nomura Yokohama City University, Yokohama, Japan

Karin Nordström Department of Neuroscience, Uppsala University, Uppsala, Sweden

Dag Nyholm Department of Neuroscience, Neurology, Uppsala University, Uppsala, Sweden

Sean I. O'Donoghue Digital Productivity Flagship, CSIRO, Sydney, NSW, Australia

Garvan Institute of Medical Research, Sydney, NSW, Australia

Tuan D. Pham The University of Aizu, Fukushima, Japan

Tatiana Polushina Faculty of Medicine and Dentistry, Department of Clinical Science, University of Bergen, Bergen, Norway

Madawa Priyadarshana Faculty of Science, Department of Statistics, Macquarie University, Sydney, NSW, Australia

Takako Sato Yokohama National University, Yokohama, Japan

Olga Senyukova Department of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russia

Seong-O Shim Faculty of Computing and Information Technology, King Abdul Aziz University, Jeddah, Kingdom of Saudi Arabia

Georgy Sofronov Faculty of Science, Department of Statistics, Macquarie University, Sydney, NSW, Australia

Christian Stolte Digital Productivity Flagship, CSIRO, Sydney, NSW, Australia

Ahmad Rauf Subhani Department of Electrical and Electronic Engineering, Centre for Intelligent Signal and Imaging Research, Universiti Teknologi PETRONAS, Tronoh, Perak, Malaysia

Changming Sun Digital Productivity Flagship, CSIRO, North Ryde, Sydney, NSW, Australia

Piotr Szul Digital Productivity Flagship, CSIRO, Sydney, NSW, Australia

Yosuke Takami Yokohama National University, Yokohama, Japan

Hai Tan Shanghai Synchrotron Radiation Facility (SSRF), Chinese Academy of Sciences, Shanghai Institute of Applied Physics, Shanghai, China

John A. Taylor Digital Productivity Flagship, CSIRO, Canberra, ACT, Australia

Darren Thompson CSIRO IM&T, Melbourne, VIC, Australia

Rafi Ullah Comsats Institute of Information Techonology, Islamabad, Pakistan

Pascal Vallotton Digital Productivity Flagship, CSIRO, Sydney, NSW, Australia

Dadong Wang Digital Productivity Flagship, CSIRO, North Ryde, Sydney, NSW, Australia

Shuaifang Wang Faculty of Health, Engineering and Sciences, University of Southern Queensland, Toowoomba, QLD, Australia

Peng (Paul) Wen Faculty of Health, Engineering and Sciences, University of Southern Queensland, Toowoomba, QLD, Australia

Tiqiao Xiao Shanghai Synchrotron Radiation Facility (SSRF), Chinese Academy of Sciences, Shanghai Institute of Applied Physics, Shanghai, China

Tomohiro Yano Yokohama National University, Yokohama, Japan

Kim Ho Yeap Faculty of Engineering and Green Technology, Department of Electronic Engineering, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia

Guohun Zhu Faculty of Health, Engineering and Sciences, University of Southern Queensland, Toowoomba, QLD, Australia

School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin, China

Acronyms

aCGH	Array comparative genomic hybridization
ALS	Amyotrophic lateral sclerosis
ANN	Artificial neural networks
ARC	ANU reference collection
ARL	Average run length
AS	Activated sludge
AU	Area under
BBME	Block based motion estimation
BMA	Block matching algorithms
BME	Block motion estimation
BOD	Biological oxygen demand
CBS	Circular binary segmentation
CE	Cross-entropy
CGH	Comparative genomic hybridization
CLSM	Confocal laser scanning microscopy
CMR	Central motor region
CMVs	Candidate motion vectors
CNV	Copy number variation
COD	Chemical oxygen demand
CT	Computed tomography
CTE	Cortical thickness estimation
CTS	Classifynder test set
CUSUM	Cumulative sum
DFA	Detrended fluctuation analysis
DO	Dissolved oxygen
DPE	Delay permutation entropy
DREAM	Dialogue for Reverse Engineering Assessments and Methods
DS	Diamond search
DT	Decision trees

ECG	Electrocardiography
EDL	Ensemble decomposition learning
EEG	Electroencephalogram
EMD	Elementary motion detector
EOG	Electrooculography
EZ	Epileptogenic zone
FBP	Filtered back-projection
FDK	Feldkamp-Davis-Kress
FFT	Fast Fourier transform
FS	Full search
FSS	Four step search
GFT	Grass fire transform
GGCM	Grey gradient co-occurrence matrix
GMM	Gaussian mixture model
GnRH	Gonadotropin-releasing hormone
HD	Huntington's disease
HMMs	Hidden Markov models
HRT	Hydraulic retention time
HRV	Heart rate variability
HS	Horizontal system
IaaS	Infrastructure as a service
IBL	Instance-based learning
ICA	Imaged absolute conics
iEEG	Intracranial EEG
KDE	Kernel density estimation
LDA	Linear discriminant analysis
LDSP	Large diamond search pattern
LH	Luteinizing hormone
LPTCs	Lobula plate tangential cells
LTR	Left temporal region
MAD	Mean absolute difference
MCRT	Mean cell residence time
MFA	Multi-fractal analysis
MGD	Million gallons per day
MIMO	Multiple-input multiple-output
MRI	Magnetic resonance imaging
MSE	Mean square error
MSEn	Multi-scale entropy
MST	Minimum spanning tree
MV	Motion vectors
NIOSH	National Institute for Occupational Safety and Health
NLD	Nonlinear dynamics
NN	Neural networks
OSA	Orthogonal series approximation
OSAlg	Orthogonal search algorithm

PaaS	Platform as a service
PAR	Peak-to-average-ratio
PD	Parkinson's disease
PDB	Protein database
PDF	Probability density function
PE	Permutation entropy
PELT	Pruned exact linear time
PET	Positron emission tomography
PRA	Pel-recursive algorithms
PSNR	Peak signal-to-noise ratio
PVC	Partial volume correction
RAS	Return activated sludge
RBF	Radial basis function
REM	Rapid eye movement
RF	Random forests
RMSE	Root mean square error
RSA	Root system architecture
RTR	Right temporal region
SaaS	Software as a service
SAD	Sum of absolute differences
SC	Stopping criterion
SDD	Sample-to-detector distance
SDI	Sludge density index
SDSP	Small diamond search pattern
SE	Sample entropy
SEL	Single-example learning
SIFT	Scale-invariant feature transform
SISO	Single-input single-output
SNR	Signal-to-noise ratio
SPEM	Smooth pursuit eye movements
SPG	Smooth pursuit gain
SPS	Smooth pursuit system
SR	Shiryaev-Roberts
SR-μCT	Synchrotron radiation based micro-computed tomography
SSIM	Structural similarity index measure
SSRF	Shanghai Synchrotron Radiation Facility
SUVR	Standard uptake value ratio
SVI	Sludge volume index
SVM	Support vector machine
TDLS	Two-dimensional logarithmic search
TDR	True detection rate
Te	Testosterone
TIE	Transport of intensity equation
TOC	Total organic carbon
TS	Total solids

TSS	Three step search
TSSol	Total suspended solids
UESA	Unimodal error surface assumption
UPDRS	Unified Parkinson's Disease Rating Scale
VG	Visibility graph
VSS	Volatile suspended solid
WWTP	Wastewater treatment plants
ZMs	Zernike moments

Part I
Signal Analysis

Chapter 1
Visual Analytics of Signalling Pathways Using Time Profiles

David K.G. Ma, Christian Stolte, Sandeep Kaur, Michael Bain, and Seán I. O'Donoghue

Abstract Data visualisation is usually a crucial first step in analysing and exploring large-scale complex data. The visualisation of proteomics time-course data on post-translational modifications presents a particular challenge that is largely unmet by existing tools and methods. To this end, we present Minardo, a novel visualisation strategy tailored for such proteomics data, in which data layout is driven by both cellular topology and temporal order. In this work, we utilised the Minardo strategy to visualise a dataset showing phosphorylation events in response to insulin. We evaluated the visualisation together with experts in diabetes and obesity, which led to new insights into the insulin response pathway. Based on this success, we outline how this layout strategy could be automated into a web-based tool for visualising a broad range of proteomics time-course data. We also discuss how the approach could be extended to include protein 3D structure information, as well as higher dimensional data, such as a range of experimental conditions. We also discuss our entry of Minardo in the international DREAM8 competition.

Keywords Visual analytics • Signalling pathways • Proteomics • Temporal data • Graph layout • Phosphorylation • Insulin response

D.K.G. Ma • S. Kaur
Garvan Institute of Medical Research, Sydney, NSW, Australia

The University of NSW, Sydney, NSW, Australia
e-mail: davidma@cse.unsw.edu.au; sandeep.kaur@unsw.edu.au

C. Stolte
Digital Productivity Flagship, CSIRO, Sydney, NSW, Australia
e-mail: christian.stolte@csiro.au

M. Bain
The University of NSW, Sydney, NSW, Australia
e-mail: mike@cse.unsw.edu.au

S.I. O'Donoghue (✉)
Digital Productivity Flagship, CSIRO, Sydney, NSW, Australia

Garvan Institute of Medical Research, Sydney, NSW, Australia
e-mail: sean@odonoghuelab.org

C. Sun et al. (eds.), *Signal and Image Analysis for Biomedical and Life Sciences*,
Advances in Experimental Medicine and Biology 823,
DOI 10.1007/978-3-319-10984-8_1

1.1 Introduction

Computationally aided data visualisation is helpful for analysing and exploring large-scale complex data as it allows computational abilities, such as large memory capacities and fast calculations, to be combined with human abilities, such as high-bandwidth visual perception and creativity, to address the task of understanding such data [19]. With the emergence of large-scale and high-dimensional datasets in molecular systems biology, the task of data visualisation has become increasingly important [28].

Current high-throughput technologies typically enable thousands of molecules to be tracked simultaneously. One such high-throughput method uses mass spectrometry to enable the quantification of the phosphorylation state of each protein in a cell's proteome. In typical experiments of this type, cells are initially stimulated with an agent (e.g., insulin, glucose, or a range of inhibitor molecules) and the response is measured at discrete points in time. The temporal order of such time-series experiments offers great potential to prioritise paths in the resulting dense protein interaction graphs [11].

In order to understand biomolecular systems it is essential to understand how the interactions of their component molecules result in the overall changes in cell physiology – for example, how a fat cell initially starved of glucose switches to active uptake and processing of glucose upon stimulation by insulin. The most common approach used to gain an understanding of such events is to draw graphs of signalling pathways [14]. These pathway maps definitely have their limitations: for example, as explained by Kitano [20], they could be thought of as analogous to static road maps, when what we really wish to know are the traffic patterns, why such patterns emerge and how we can control them. Nonetheless, visualisations of pathway maps are an important first step.

There are several initiatives worldwide aimed at consolidating all human knowledge about biological systems into a single, searchable database and with the results presented in the form of interactive pathways graphs. Currently however there is no consensus about a single 'best' approach – instead, there are a large number of different databases, each with a tailored visualisation system. Some of the more widely used resources of this kind include Pathway Commons [6], KEGG [18], PANTHER [25], BIOCARTA [27], and Reactome [24].

When considering new data from high-throughput experiments, a common strategy is to visually overlay these data onto existing pathway graphs extracted from one of the above resources. A wide variety of methods and tools have been developed to facilitate overlaying experimental data onto pathways, including Pajek [3], BiologicalNetworks [1], Medusa [31], as well as many plug-ins to the Cytoscape framework [36]. A recent review of such methods for 'omics' data is provided by Gehlenborg et al. [14]. However, as noted in [14], the major challenge for visualisation methods is how to benefit from the explosion in dataset scale and complexity without overwhelming the user. This is a difficult problem which

currently has no obvious general solution, but we suggest the answer should lie in how *context* may be used in visualisation. The contribution of this paper lies in the adoption of a novel visual metaphor that can illustrate significant temporal and potentially causal relationships in high-throughput data on cell signalling pathways.

1.1.1 Challenges in Visualising High-Throughput Time-Series Post-translationally Modified Proteomic Datasets

The biology of the post-translational modification of proteins presents some important issues for visualisation. Firstly, there are many different types of such modifications that we may require to be visualised (e.g., phosphorylation, methylation, sumoylation, etc.). Since different modifications are typically implicated in different functional roles, indications of these differences could be critical for successful visualisation.

Secondly, most of the current network-based visualisation tools for high-throughput datasets have been designed for gene expression. However, it is not always possible to simply reuse such tools for proteomics datasets that incorporate post-translational modifications. For example, when viewing time-series transcriptomic datasets, we are usually interested in the expression levels of *whole RNA molecules* over time – for such data, time-profile information is often added to a network view by adding colouring or a pattern to each node or edge [14, 35]. However for proteomics datasets showing post-translational modifications a more detailed representation is required, since we are typically interested in the abundance levels of *multiple residues* within each protein.

As an example, we recently conducted a pilot user study [21] to evaluate the reuse of the Cytoscape plug-in 'Cerebral' [2] to visualise the proteomic dataset of Humphrey et al. [15]. Cerebral was initially designed for use with gene expression data – we found that although several aspects of this tool were of benefit, overall the layout and representation concepts of the tool were not well suited to visualising post-translational modifications.

A general problem faced by omics visualisation tools is the challenge of facilitating simultaneous visualisation of multiple kinds of experimental data. For example, how can high-throughput time-series data on post-translational modifications be visualised in an coherent and integrated way with other data, such as abundance level for transcripts or protein?

Evidence of the growing recognition of the need and importance of this type of integrated omics visualisation comes in the form of the latest iteration of the international DREAM (Dialogue for Reverse Engineering Assessments and Methods) competition. The DREAM8 Sub-challenge 3: "Visualisation of high-dimensional time-course on breast cancer proteomics data" was designed to facilitate research on novel tools for this purpose.

1.1.2 Aims

In this paper we outline the elements of the Minardo visualisation concept (Fig. 1.1) that we developed recently to address the above challenges – Minardo is based on using cell topology combined with temporal ordering as the key layout contexts used to organize how the data is depicted. In this study, we worked with an experimental research group that is applying state-of-the-art methods in high-throughput experimental proteomics to study the time course of protein phosphorylation events in human cells in vitro following stimulation by insulin [15], as part of a broader project on diabetes and obesity. The group had already applied a wide range of existing analysis and visualisation tools to these data, although relatively few tools were specifically tailored for time-course phosphophorylation data. The group's key unmet requirement was for a system that would enable visual exploration of networks representing insulin response, which could be interactively overlaid firstly with phosphophorylation time-course data, and that later could also include data on RNA and protein abundance.

To address this need, we first tried several existing visual analytics approaches with the goal of representing the data to gain new insight. From discussion of the merits and weaknesses of these existing approaches with our experimental collaborators we used this feedback, together with visual analytics principles, to develop an improved general layout strategy specifically for time-series post-translationally modified proteomic data.

We called our layout strategy "Minardo" as a play on words, as the layout was partly inspired by the well-known information graphic published by Minard in 1869.[1] A key innovation that comes from this inspiration is the ability to combine aspects of network-based structure with temporally ordered event profiles.

As discussed in [21], the Minardo approach has proved helpful, having revealed several inconsistencies with the previously 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 has been constructed specifically for analysing phosphorylation data related to insulin response, aspects of the layout have clear potential to be generalised to help with analysing a broader range of systems biology data. Thus, we are doing ongoing work aimed at developing the Minardo layout strategy into a general tool.

The remainder of this paper is structured as follows. In Sect. 1.2 we describe how to create the Minardo layout and how to connect it interactively with a heat map visualisation of the same data. Section 1.3 presents results from a user study to evaluate Minardo, and our entry using Minardo into the DREAM8 visualisation challenge. In Sect. 1.4 we discuss implications of this approach and outline directions for future work. Section "Conclusions" concludes the paper.

[1]This famous graphic shows Napoleon's disastrous Russian campaign of 1812 – the graphic is regarded as an exemplar by many data visualisation specialists [4].

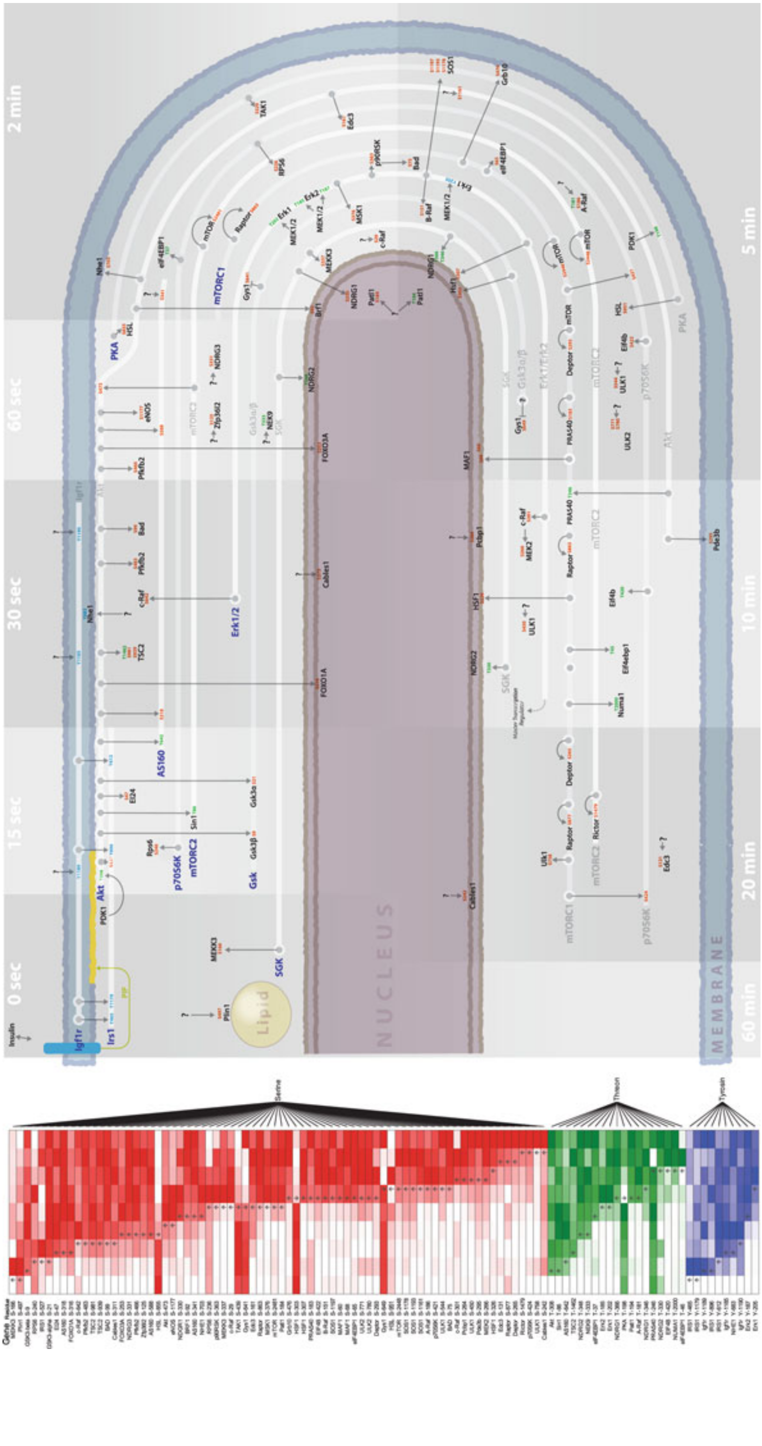

Fig. 1.1 Screenshot of the interactive heat map and Minardo layout for the insulin response dataset [15] (Reprinted with permission from [21]. Copyright 2013, AIP Publishing LLC)

1.2 Methods

Our visualisation strategy consists of two main components – the Minardo layout, and a heat map – both of which are connected to support interactive data exploration. This is shown in Fig. 1.1, using the insulin response dataset of Humphrey et al. [15].

The Minardo layout depicts a cellular topology, divided into regions that represent the time points of the time course data. Note that although time points typically denote discrete values fixed by the experimental protocol (e.g., 0, 15, 30 s, etc.), the Minardo layout allows placement of events ordered along a continuous time scale in cases where continuous data are available (e.g., either directly from experiments that measure continuous time values, or perhaps estimated by interpolation from discrete data). The tracks across time points indicate individual proteins or protein complexes that are active (in terms of events in the dataset) over multiple time points. Since screen real estate is limited, only a limited number of tracks can be displayed at the same time. In many cases the number of tracks that can be shown will be insufficient to show all proteins active across multiple time points – thus, some criteria will need to be applied to select which proteins are to be displayed on the available tracks. A natural criterion is the level of activity of a protein, as this suggests its importance, although other criteria could also be applied.

Causal relationships between different proteins within a time point are depicted using directed edges running perpendicular to the tracks; for the insulin response dataset [15], these relationships link a kinase to its phosphorylation substrate. The actual protein residue number of the phosphorylated amino acid – known as the *phosphosite* – is shown colour-coded in Fig. 1.1.

In the rest of this section we describe in further detail the method and the design decisions used to create the layout. We also describe the heat map, and the procedures used to link the layout to the heat map, as well as the implementation procedures used to make the visualisation interactive.

1.2.1 Phosphorylation Dataset for Insulin Response

We worked with members of the James laboratory[2] at the Garvan Institute of Medical Research, a world-leading laboratory in applying experimental systems biology to study diabetes and obesity. They recently published a study of the time course of protein phosphorylation events occurring in vitro in mouse 3T3-L1 adipocyte cells – cells derived from brown adipose (fat) tissue – after 'feeding' (i.e., stimulation by insulin and glucose) [15].

The cell used in the experiment was initially in a starved state, then stimulated with insulin and glucose. The cells were lysed at 0, 15, 30 s and 1, 2, 5, 10, 20 and 60 min after stimulation. Mass spectrometry was then used to measure the

[2] http://www.jameslab.com.au

phosphorylation state of all detectable Serine (S), Threonine (T) and Tyrosine (Y) amino acid residues [9], 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 [15].

Humphrey et al. then used unsupervised fuzzy C-means clustering to organise the time profiles for each phosphosite into groups [15]. They also conducted an extensive literature survey to identify the kinases responsible for a subset consisting of 104 of the phosphosites judged to be most significant, based on prior knowledge of the response pathway. These data – presented in Fig. 5 of Humphrey et al. [15] – were used as the starting point for our work, with the goal of re-analysing and organising these data to provide greater insight into underlying biological processes.

1.2.1.1 Data Representation

The phosphorylation time-series data from Humphrey et al. [15] were generated by the MaxQuant software [10]. We obtained these data as a comma-separated text file, which contained the ratios of the absolute values of observed levels of phosphorylation for each phosphosite at each time-point to a basal level. The basal level represented the phosphorylation levels in starved cells, and the time-points represented the phosphorylation levels after that amount of time has elapsed since stimulation with insulin and glucose. The dataset consisted of triplicate measurements of phosphorylation levels for each of the nine time-points. Phosphorylation levels at time zero were set to 1.0, and the phosphorylation level at each subsequent time point was the ratio of that point's abundance to its basal level (for more information, see the Methods section of Humphrey et al. [15]).

1.2.2 Heat Map of the Time-Series Data

To display the complete time-course data, we used a traditional heat map. We utilised three colour scales, red, green, and blue to represent *Serine*, *Threonine* and *Tyrosine* residue phosphosites, respectively. The heat map depicted only those profiles which were also present in the Minardo layout.

In order to create the heat map, we averaged the triplicate values at each time-point, then we linearly rescaled the resulting time profile, setting the lowest level of activation achieved across the time-series to 0 % and the highest level to 100 %. Finally, we used the JavaScript library D3.js [5] to create an interactive heat map visualisation of these data.

1.2.2.1 Selecting a Single Time Point for Each Phosphorylation

Using the re-scaled data, we devised a method for consistently selecting a single representative time-point for each phosphosite. Based on an analogy to the Michaelis

constant [26] in enzyme kinetics, we estimated the time at which each phosphosite first transitions from either below its 50 % level to above, or vice versa in the case of a dephosphorylation event. We took this to be the first activation time, and marked it on the heat map using either an up or down arrow to indicate phosphorylation or dephosphorylation, respectively.

This induces a linear (total) ordering on the data, where each phosphosite's time course is denoted by its estimated first activation time. Note also that the activation times estimated by this method have continuous values, hence effectively increasing the temporal resolution of the dataset, which is also useful for constructing the Minardo layout.

1.2.3 The Minardo Layout

The Minardo layout was constructed using a number of graphic design principles, combined with user feedback, and drawing from concepts used in existing tools, such as the Cerebral plug-in [2]. The visual channels used – primarily position, hue, and connection – were chosen to effectively convey key information with low cognitive load [7, 38].

Position is usually the most powerful visual channel [22] hence in Minardo we have used the X and Y axis to show time and sub-cellular topology, as they are key features of the dataset. We created a schematic cell in Adobe Illustrator, mapping time in an arc around the cell and adding intervals to represent the time points used to derive the experimental data (Fig. 1.1). With a single first activation time identified for each phosphosite, phosphorylation events could be placed unambiguously within one specific time interval on the diagram. We also arranged the cell topology such that the regions for each time interval contain extracellular space, cytoplasm, and nuclear space, allowing for positioning proteins based on their subcellular location.

Rather than laying out the consecutive time intervals in one direction (e.g., along the X or Y axes), we have taken inspiration from Charles Joseph Minard's classic flow map of Napoleon's March [4] and wrapped the flow of time around the cell topology, creating an overall aspect ratio that allows the entire diagram to more easily fit the landscape orientation of most computer displays. Wrapping time in this way also allow connections from later to earlier time points, providing clear representation of feedback loops.

Lines with arrows were used to indicate kinases and their target phosphosites. In the current dataset there are 104 such connections. To overcome the typical 'hairball' problem that occurs with networks of this size and larger, we reduced clutter by using tracks to represent 'promiscuous' proteins or complexes, i.e., those involved in multiple phosphorylation events at 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-series dataset.

Hue was used consistently in the network and heat map, with red, green, and blue used to represent Serine, Threonine and Tyrosine residues. Yellow was used to highlight items selected by the user. The default highlight was Yellow only,

and it showed the relevant kinases and phosphorylation events on the track, or the phosphates currently being brushed over. "Show Targets" is a toggle button, which turns Teal when switched on, indicating to the user that phospho-targets are now being shown with a Teal highlight.

The layout was saved in SVG format and imported in an HTML page with the heat map. JavaScript was used to implement brushing and linking between the two representations.

1.3 Results

Our implementation of the Minardo layout and heat map applied to the insulin response phosphorylation time course data [15] resulted in a single HTML file, which is included in the supplementary information (http://odonoghuelab.org/Minardo.zip). A screen shot of this HTML file can be seen in Fig. 1.1. Two distinct components are clearly seen in this figure, the heat map and the Minardo Layout. The HTML files supports interactivity between these components via brushing and linking. For example, hovering over a protein (in either the heat map or the Minardo layout) automatically highlights all occurrences of the protein name in the HTML document. Text searching of the HTML document can also be done, using standard browser functionality, resulting in highlighting of all proteins with names that match the search term.

In the Minardo layout, the insulin response network (taken from Fig. 5 of Humphrey et al. [15]) has been overlaid on a typical cellular topology. This cellular topology has been divided into a number of time-points as present in the dataset – in this case, nine time points. It shows the temporal order of phosphorylation events, with arrows identifying each kinase and, its substrate phosphosite. The proteins Akt, Irs1, AS160, p70S6K, Erk1 and Erk2, and the complexes Gsk, mTORC1 and mTORC2, play roles across multiple times and so have been indicated with white tracks running parallel to the cellular membrane. For each of the protein phosphorylation sites featured in the Minardo layout, an entry has been created in the interactive heat map, showing its normalised abundance levels detected across each of the time points. The HTML allows sorting the heat map in multiple ways including by residue type, by UniProt identifier [23], by identified time of first regulation, and many more.

1.3.1 Evaluation of the Minardo Visualisation Strategy

We conducted an informal user study with experts in the field of diabetes and obesity studying insulin response at the Garvan Institute [17]. During the study, the interactive HTML file was disseminated to the users by making it available within the organisation's intranet. Users were asked to freely explore the use of the visualisation strategy and provide feedback.

To summarise the results of this study, the users judged the Minardo visualisation favourably. They found the brushing and linking feature between the heat map and the network to be very helpful for interpreting the data in detail. The most positive feedback, however, was that the new layout helped them gain new insights into the underlying bio-molecular processes. These new insights are detailed in our related work in Ma et al. [21]. The validations of these insights are underway, with a joint publication with the biologists in preparation.

1.3.1.1 Requested Features

In this study, the users requested a number of features not yet supported by our current Minardo implementation.

First and foremost was the ability to easily select which phosphorylation sites are used to construct the layout. The current dataset shows only 104 of the 7,897 high quality phosphorylation sites that were present in the original dataset – these 104 sites were selected by Humphrey et al. [15] as they were believed to be the most important. Nonetheless, the users would like the facility to examine other subsets of phosphosites using the Minardo layout.

A second requested feature was the ability to interactively edit the network in order to change the assignment of kinases to targets. A third requested feature was the ability to add additional data (or datasets), such as multiple experimental conditions or the presence of various chemical inhibitors. Finally, users requested that the facility to search proteins by name be extended so as to match different synonyms for the same protein – this would be very useful since many proteins used in this study are known by multiple names (e.g., As160, Kiaa0603, Tbc1d4 all refer to the same protein).

1.3.2 Minardo in the International DREAM8 Competition

The aim of the DREAM8 'visualisation of high-dimensional time-course on breast cancer proteomics data' sub-challenge was to propose novel strategies to visualise high-dimensional molecular time-course data. The datasets provided for the competition featured phosphorylation proteomics data, for approximately 45 phosphosites, at seven time points, under eight stimulus conditions. Data was also given for a control, and under conditions in which the phosphorylation ability of 3 crucial kinases was inhibited.

We entered Minardo in this competition. All of its features as described above were presented, however additional modifications were proposed to its workflow to enable comparison between multiple different experimental conditions (stimuli, inhibitor or control). Figures 1.2 and 1.3 show the main visualisations in the proposed Minardo workflow when visualising a single set of conditions and when comparing between two conditions.

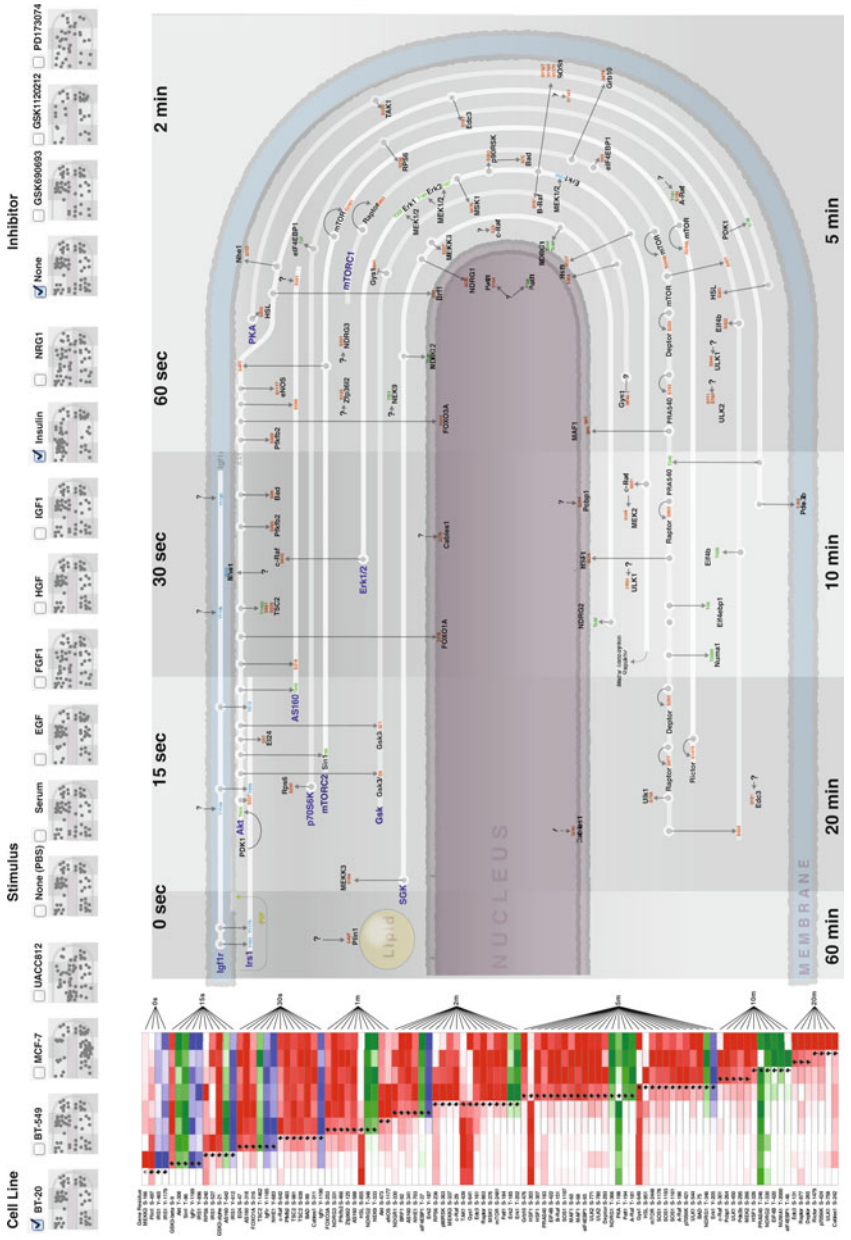

Fig. 1.2 Visualising a single set of conditions as selected interactively by the user. Additional checkboxes are placed above the Minardo layout to enable the user to select from a set of conditions (in this case, to pick a Cell Line, a Stimulus, and Inhibitor). The rest of the visualisation is as before