



Data-Driven Modelling and Optimised Reverse Engineering of Complex Dynamical Systems in Cancer Research

Michael Adewunmi Idowu

Submitted to the University of Abertay
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

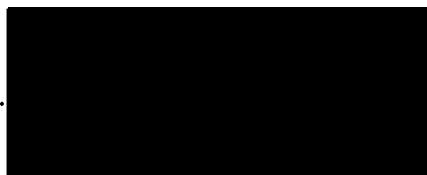
July 2013

© Michael A. Idowu 2013.

All rights reserved.

I certify that this thesis is the true and accurate version of the thesis
approved by the examiners.

Signed.....

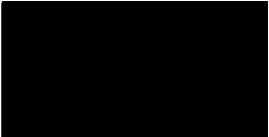


.....
(Director of Studies)

Date.....25/11/13.....

Declaration

I, Michael Adewunmi Idowu, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I can confirm that this has been indicated in the thesis.

Date ... 

Data-Driven Modelling and Optimised Reverse Engineering of Complex Dynamical Systems in Cancer Research

by

Michael Adewunmi Idowu

Submitted to the University of Abertay
on July 1, 2013, in partial fulfilment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Biological systems typically generate complex data that encapsulate the dynamics of interactions among measurables over time. To support the formation of insights into time series data from a biological system, there is a requirement to develop new methods that can analyse and translate such complex data into a form that allows trends, patterns, and predictions to be easily viewed, verified and tested. Here, a suite of novel analytical and matrix-based techniques for dynamical systems modelling are developed that are time-efficient and data-driven. These techniques facilitate a range of scientific analyses through novel matrix-based system identification and parameter estimation methods. The inference techniques are fast, optimised, and do not require *a priori* information to successfully infer network of interactions or automatically construct data-consistent models from data. Two distinct principal (Jacobian and power-law) models (solutions) that are data-consistent may be constructed from a single time series data set. A recast technique has also been developed to reconstruct either one of the principal models from the other, providing support for model interoperability and multiple model integration.

The thesis demonstrates the effectiveness of a new theoretical framework developed to incorporate a modelling and visualization pipeline able to deal with a wide range of time-series data sets relating to complex biological systems. The integrated framework is able to infer and depict interaction networks implicit in time series data in just a matter of seconds and then display the evolution of that network dynamics in response to network perturbation such as drug treatments. Beyond this, there is a broader contribution to the field of biochemical system theory (BST), evidenced by establishing methods for transforming a constructed jacobian model to equivalent power-law models, and vice versa. The effectiveness of these new techniques is demonstrated using artificial time series data samples, simulated pseudo-data of biologically plausible models of real biological systems, and real experimental data derived from biological experiments.

Acknowledgements

First and foremost, I would like to thank the Almighty God, the Father of my Lord and Saviour Jesus Christ, without whom all efforts would have been in vain. To You Father be all the glory, honour and adoration both now and forever more. Thank you for the divine inspiration, wisdom and understanding that always come through the power of Your Holy Spirit.

I also would like to thank the following people for their contribution to my success in completing this thesis: Prof. James Bown for being a great supervisor and mentor; Prof. Nikolai Zhelev for his assistance, funny jokes and good advice; Prof. John Crawford for being the first to introduce me to mathematical modelling.

I dedicate Chapter 5 to Dr. Hilal Khalil, a colleague and true friend whose insightful comments have always been very helpful. Dr. Alexey Goltsov deserves to be acknowledged for his contribution to Chapter 6 for providing time series data: thanks Alexey, for the process-based simulation and interesting discussions.

Many thanks to the examiners, Dr Ruth Falconer, Dr Janine Illian and Dr Simon Langdon, for their constructive and encouraging feedback.

To the Abertay Research Committee, thank you very much for your support and assistance. To Prof. John Palfreyman, Dr Nia White and Diane Norris, thank you for your advice and patience. To all members of SIMBIOS, School of Contemporary Sciences, School of Computing and Creative Technologies, Information Services, Estates and Campus Services, thank you very much.

To my adorable wife Shade, most caring parents, and lovely children Toluwani, 'Tobi, 'Tosin and Joshua: thank you all for your support, patience and understanding. I am most grateful for the love and sacrifices.

List of Publications

1. Idowu M.A.; Bown J.L. Towards an exact reconstruction of a time-invariant model from time series data *Journal of Computer Science and Systems Biology* Vol. 4 (4): 055-070 (2011) - 055. DOI: 10.4172/jcsb.1000077, 2011.
2. Idowu M.A.; Bown J.L. Matrix Operations for the Simulation and Immediate Reverse-Engineering of Time Series Data UKSim 14th International Conference on Computer Modelling and Simulation, DOI: 10.1109/UKSim.2012.24 Page(s): 101 - 106, 2012.
3. Idowu M.A.; Bown J.L. Bown, Matrix-Based Analytical Methods for Recasting Jacobian Models to Power-Law Models. 8th EUROSIM Congress on Modelling and Simulation, Computer Modelling and Simulation (EuroSim), Digital Object Identifier: 10.1109/EUROSIM.2013.53 Page(s): 250 - 258, 2013.
4. Idowu M.A., "Improved Modelling of Dynamic Systems", International patent application: WIPO WO2013/024293; PCT/GB2012/051997.
5. Bown J.; Andrews P.S.; Deeni Y.; Goltsov A.; Idowu M.; Polack F.A.; Sampson A.T.; Shovman M.; Stepney S. Engineering simulations for cancer systems biology. *Curr Drug Targets*. 13(12):1560-74, 2012.
6. Goltsov A.; Deeni Y.; Khalil H.; Idowu M.; Kyriakidis S.; Goltsov G.; Langdon S.; Harrison D., Bown J. Role of Post-translational Regulation of PTEN Activity in Cancer Cell Addiction to Heterozygous PTEN Mutations. Pages 173-210; ISBN: 978-1-62808-049-0, 2013.
7. Idowu M.A.; Bown J.; Zhelev N. A new method for identifying a data-consistent self-reconfigurable predictive bio-network model of the cell cycle based on time series data and its application in cancer systems biology. Annual meeting, American Association for Cancer Research, 2012.
8. Idowu M.; Goltsov A.; Khalil H.S.; Tummala H.; Zhelev N.; Bown J. Cancer research and personalised medicine: a new approach to modelling time-series data using analytical methods and Half systems, *Current Opinion in Biotechnology*, Volume 22, Supplement 1, Page S59, ISSN 0958-1669, 10.1016/j.copbio.2011.05.163, 2011.

9. Shovman M.; Idowu M.; Goltsov A.; Bown J. Dynamic visualisation of biological network models, Intl. Conf. on Systems Biology, Edinburgh, 2010.

Contents

1	Introduction	19
1.1	Motivation	25
2	Cancer biology and mathematical modelling	27
2.1	Causes of cancer	27
2.2	Background on the essentials of cancer biology	29
2.2.1	Growth factors and the HER family of receptors	30
2.2.2	HER receptor ligand-binding	31
2.2.3	Conformational change	31
2.2.4	Receptor dimerization (pairing)	32
2.2.5	Cancer and HER receptor dysregulation	32
2.2.6	Intracellular signalling pathways	33
2.2.7	Diagnosis and treatment of breast cancer	34
2.3	Cancer systems biology and computational Modelling	36
2.4	Process-based modelling	41
2.4.1	Major challenges of process-based modelling	42
2.5	Data-driven modelling	43
3	System identification methods	47
3.1	Data-driven Modelling Approach	47
3.1.1	ODE based modelling	48
3.1.2	Systems representation, identification and parameter estimation	48
3.2	A review of deterministic modelling approaches	52
3.3	Understanding BST as a modelling approach	61
3.3.1	Analytical convenience of BST	61

3.3.2	Half system ODE representation	62
3.3.3	S-system ODE representation	63
3.3.4	The Generalised Mass Action (GMA) System	64
3.4	Modelling based on BST	65
3.4.1	System identification and parameter estimation considerations . . .	66
3.4.2	Major challenges of BST model and methods	68
4	Novel reverse engineering and network inference methods	71
4.1	Method 1: core method - jacobian approach	71
4.2	Ordinary differential equation systems	73
4.2.1	Problem statement	74
4.2.2	Relative rate of change	74
4.2.3	Complex nonlinearity in systems of nonlinear ODEs	76
4.2.4	Methodology	77
4.2.5	Transposive and repressive regression methods	81
4.2.6	The search for the jacobian matrix solution	83
4.2.7	Application of eigenvalues and eigenvectors	83
4.2.8	A new method for calculating matrix logarithmic inverse	84
4.2.9	Linking jacobian matrices and network models	84
4.2.10	Results	85
4.2.11	Method validation	85
4.2.12	Data discretisation using a simple continuous model	86
4.2.13	Data discretisation using eigenvectors and eigenvalues	88
4.2.14	Results: application of reverse engineering methods	93
4.2.15	Performance evaluation of algorithms	96
4.2.16	Algorithm performance	98
4.2.17	Conclusion: jacobian method	101
5	Dynamic modelling of DNA-damage response (DDR) pathways	103
5.1	Understanding the DNA-damage response pathway	104
5.2	Aims and objectives	105
5.2.1	DDR cancer biology	106
5.2.2	Biological experiments and method	108

5.3	Dynamic modelling and system identification methods	109
5.3.1	DDR modelling challenge	110
5.3.2	Computational modelling objectives	110
5.3.3	An overview of the computational modelling approach	111
5.3.4	Application of modelling methods	113
5.3.5	Use of heatmaps	114
5.4	Analysis and interpretation of results	115
5.4.1	The constructed jacobian and Half-system models	115
5.4.2	Initial analysis of data	116
5.4.3	Interpretation of results	123
5.4.4	Further analyses of segments of experimental data	126
5.5	Discussion and conclusion	129
6	Dynamic modelling of PI3K-AKT signalling pathways	133
6.1	Background	138
6.2	Understanding signal transduction	139
6.2.1	Major modelling challenges	140
6.2.2	Towards multiple parameter fits	141
6.3	Reverse engineering of RTK-PI3K-MAPK signalling pathways	142
6.3.1	Problem definition	142
6.3.2	The datasource model of input data: the PI3K / PTEN / AKT sig- nalling networks	143
6.3.3	The acquired data samples	145
6.3.4	Pertuzumab (2C4): a monoclonal antibody that targets the HER family of cell-surface receptors	146
6.3.5	Modelling of <i>in-silico</i> experimental time series data	148
6.4	Presentation of modelling and inference results	155
6.4.1	General consideration	155
6.4.2	Result and interpretation	155
6.5	Remarks	157
7	Conclusions	161
7.1	Confirmation of hypotheses	161

7.1.1	Hypothesis 1	162
7.1.2	Hypothesis 2	163
7.1.3	Hypothesis 3	164
7.1.4	Hypothesis 4	165
7.2	Concluding remarks	166
7.3	Future work and considerations	168
A	Figures	171
B	Figures	175
C	Tables	177
D	Figures	181
E	Supplementary information	189
E.0.1	Calculate E_1	190
E.0.2	Calculating EE_1 : an alternative method	191
E.0.3	Assessment of initial results	194
E.0.4	Major challenge	195
E.0.5	Finding a superlative (jacobian) solution	196
F	Alternative methods: matrix-based analytical techniques	199
F.1	Method 2: heuristic development of new analytical methods	199
F.1.1	Multiple model integration	200
F.1.2	Extending biochemical system theory (BST) framework	201
F.2	Method 3: Half-system based inference algorithm	202
F.2.1	Half-system: estimation of kinetic parameters	204
F.2.2	Relating the jacobian to half-system model	205
F.2.3	Estimating fractions of kinetic parameters in pairs	206
F.2.4	Validating the calculated kinetic orders	208
F.2.5	Vectorisation of estimated ratios of kinetic orders (parameters)	209
F.2.6	Matriculation of estimated ratios of kinetic orders (parameters)	209
F.2.7	Inverse diagonalisation of the principal entries of the jacobian	210
F.2.8	Derivation of the kinetic orders matrix of the half-system model	212

F.2.9	Relation between the jacobian and kinetic orders matrix	212
F.2.10	Significant contribution to BST: new recast technique (BAE)	213
F.2.11	Application of new recast method to real experimental data	215
F.2.12	Conclusion: power-law method	216
G	New matrix construction and decomposition methods	219
G.1	A new matrix decomposition and composition method	221
G.1.1	Definitions	221
G.1.2	Representation of matrix entries by minors	223
G.1.3	Construction of 4x4 matrices	223
G.2	Matrix construction with fixed determinant(s)	224
G.2.1	Example #1: matrix composition	224
G.2.2	Example #2: matrix composition with fewer parameters	225
G.2.3	Example #3: symmetric matrix composition	225
G.2.4	Example #4: variant symmetric matrix composition	226
G.2.5	Multiple matrices with a predefined (fixed) determinant	226
G.3	Decomposition of matrices	226
G.3.1	$L_d.D_d.U_d$ Decomposition of a Symmetric Matrix	226
G.3.2	Relation between Cholesky and $L_d.D_d.U_d$ decomposition methods	227
G.3.3	Relation between LU and $L_d.D_d.U_d$ factorisation methods	228
G.3.4	Other variants of our $L_d.D_d.U_d$ decomposition method	230
G.4	Applications of $L_d.D_d.U_d$ method to systems of linear equations	230
G.4.1	Solving systems of linear systems	230
G.4.2	Application to time series inverse problem analysis	231
G.4.3	Solving time series inverse problem using matrix manipulation	233
G.5	Conclusion: matrix construction and decomposition method	233

List of Figures

1-1	Thesis structure.	25
2-1	A proposed robust and inexpensive matrix-based reverse engineering framework that is able to optimally utilise limited time series data.	44
4-1	Relation between the derived boolean representation of the jacobian matrix and the corresponding network topology with inter-connected nodes.	97
5-1	Understanding cellular responses to DNA-damage response pathways and ATM as a mediator of responses to DNA-damage.	106
5-2	Time series measurements at lower ($0.1\mu\text{M}$) and higher ($0.4\mu\text{M}$) dose-intensities of both doxorubicin and doxorubicin+KU treatments.	112
5-3	A new dynamic modelling and reverse engineering strategy for analysing time series data of DNA damage response pathway to infer and construct a data-consistent predictive model of the system.	114
5-4	The reverse engineered jacobian models that are consistent with historical time series measurements at lower ($0.1\mu\text{M}$) and higher ($0.4\mu\text{M}$) dose-intensities of doxorubicin with and without KU treatment.	118
5-5	Simulation of system dynamics: consistent with historical time series measurements at ($0.1\mu\text{M}$) dose-intensities of doxorubicin without KU treatment.	119
5-6	Simulation of system dynamics: consistent with historical time series measurements at ($0.4\mu\text{M}$) dose-intensities of doxorubicin without KU treatment.	119
5-7	Simulation of system dynamics: almost consistent with historical time series measurements at ($0.1\mu\text{M}$) dose-intensities of doxorubicin with KU treatment.	120
5-8	Simulation of system dynamics: consistent with historical time series measurements at ($0.4\mu\text{M}$) dose-intensities of doxorubicin with KU treatment.	120

5-9	Derived topological map of network of DDR signalling pathway at $0.1\mu\text{M}$ Dox input in the absence of ATM inhibition.	121
5-10	Derived topological map of network of DDR signalling pathway at $0.4\mu\text{M}$ Dox input in the absence of ATM inhibition.	121
5-11	Derived topological map of network of DDR signalling pathway with combinatorial treatment at $0.1\mu\text{M}$ Dox and with ATM inhibitor.	122
5-12	Derived topological map of network of DDR signalling pathway at $0.4\mu\text{M}$ Dox input in the presence of ATM inhibition.	122
5-13	Further analysis: reverse engineered jacobian models that are consistent with historical time series measurements at ($0.4\mu\text{M}$) dose-intensities of doxorubicin with and without KU treatment using data with timepoints 0-8hr.	128
5-14	Further analysis: reverse engineered jacobian models that are consistent with historical time series measurements at ($0.4\mu\text{M}$) dose-intensities of doxorubicin with and without KU treatment using data with timepoints 12-24hr.	128
6-1	Schematic representation of the HER 2/3-PI3K-MAPK signalling pathways.	137
6-2	Schema of process-based model of RAF/MEK/ERK and PI3K/PTEN/AKT signaling network.	145
6-3	Result of S- normalised data, 8 minutes.	151
6-4	Result of S- normalised data, 10 minutes.	151
6-5	Result of S- normalised data, 12 minutes.	151
6-6	Result of R- normalised data, 8 minutes.	152
6-7	Result of R- normalised data, 10 minutes.	152
6-8	Result of R- normalised data, 12 minutes.	152
6-9	Result of S+ normalised data, 8 minutes.	153
6-10	Result of S+ normalised data, 10 minutes.	153
6-11	Result of S+ normalised data, 12 minutes.	153
6-12	Result of R+ normalised data, 8 minutes.	154
6-13	Result of R+ normalised data, 10 minutes.	154
6-14	Result of R+ normalised data, 12 minutes.	154
A-1	The initial proposed system identification (inference) framework.	172
A-2	Evaluation of optimum results (100 networks).	172

A-3	Evaluation of optimum results (50 networks).	173
A-4	Assessment and comparison of optimal system identification methods	173
A-5	Assessment and comparison of optimal system identification methods	174
A-6	Optimum result selection: “best overestimates” and “best underestimates” .	174
B-1	The equivalent derived half-system representations of the four systems. . .	176
D-1	Heat map representations of 4 different normalised data sets generated from a biologically plausible process-based model of HER2/3-PI3K-MAPK sig- nalling pathways with 8-timepoint readings recorded over a period of 8 min- utes only.	182
D-2	Heat map representations of 4 different normalised data sets generated from a biologically plausible process-based model of HER2/3-PI3K-MAPK sig- nalling pathways with 8-timepoint readings recorded over a period of 10 minutes only.	182
D-3	Heat map representations of 4 different normalised data sets generated from a biologically plausible process-based model of HER2/3-PI3K-MAPK sig- nalling pathways with 8-timepoint readings recorded over a period of 12 minutes only.	183
D-4	Using signal transduction network data to infer or predict signalling from enzyme-coupled cell-surface receptors to intracellular kinases: modelling a) drug resistance (L.H.S) and b) sensitivity to RTK inhibition (R.H.S) purely from data generated from well-tested process-based models that switched between these two modes.	183
D-5	Derived heatmap of signalling network of HER 2/3-MAPK/PI3K signalling pathways obtained through network inference method applied on normalised data (with only 8 time points).	184
D-6	Derived heatmap of signalling network of HER 2/3-MAPK/PI3K signalling pathways obtained through network inference method applied on normalised data (with only 10 time points).	185
D-7	Derived heatmap of signalling network of HER 2/3-MAPK/PI3K signalling pathways obtained through network inference method applied on normalised data (with only 12 time points).	186

D-8	Corresponding result of S-, S+, R-, R+ data, absolute data, 8 minutes.	. .	187
D-9	Corresponding result of S-, S+, R-, R+ data, absolute data, 10 minutes.	.	187
D-10	Corresponding result of S-, S+, R-, R+ data, absolute data, 12 minutes.	.	188

List of Tables

3.1	Summary of ODE-based deterministic modelling methods	60
4.1	Reverse engineering method: pseudo code	94
4.2	Summary statistics of algorithm performance	98
5.1	Treatment conditions for Neutral Red (NR) uptake based cell cytotoxicity assay.	109
5.2	DDR substrates analysed in the study	109
6.1	Analyses of topological maps data.	157
C.1	S-, normalised data, 8 minutes. Inferred matrix of signalling network	177
C.2	S+, normalised data, 8 minutes. Inferred matrix of signalling network . . .	177
C.3	R-, normalised data, 8 minutes. Inferred matrix of signalling network	177
C.4	R+, normalised data, 8 minutes. Inferred matrix of signalling network . . .	178
C.5	S-, normalised data, 10 minutes. Inferred matrix of signalling network . . .	178
C.6	S+, normalised data, 10 minutes. Inferred matrix of signalling network . .	178
C.7	R-, normalised data, 10 minutes. Inferred matrix of signalling network . . .	178
C.8	R+, normalised data, 10 minutes. Inferred matrix of signalling network . .	178
C.9	S-, normalised data, 12 minutes. Inferred matrix of signalling network . . .	179
C.10	S+, normalised data, 12 minutes. Inferred matrix of signalling network . . .	179
C.11	R-, normalised data, 12 minutes. Inferred matrix of signalling network . .	179
C.12	R+, normalised data, 12 minutes. Inferred matrix of signalling network . .	179

Chapter 1

Introduction

Drug development and clinical testing is recognised to be a time-consuming and expensive process (DiMasi J.A. 2003) with investors seeking maximum returns on minimum investment. In today's global economy, the pharmaceutical sector must compete economically with other commercial sectors, while at the same time overcoming the sector-specific concern of a substantial translational gap between biomedical funding and results, i.e. new drugs (Dorsey E.R. 2009). Additionally, there is the humanitarian goal to develop new, inexpensive and life-saving drugs that can treat patients quickly and efficiently (ImpactReport 2002, FDA 2012). Methods that help reduce development time and direct the discovery process - so reducing costs - can contribute to the competitiveness and humanitarian value of the pharmaceutical industry.

Rapid advances in the design and development of high-throughput technologies and methods that are capable of generating large amounts of data demand new mathematical approaches that can manage or handle routine analysis and modelling tasks in faster time. This requires an establishment of a sound theoretical and model-based approach that can cope with contemporary modelling challenges, including modelling of both limited and large quantities of time series data.

With respect to understanding complex systems, mathematical modelling may be employed to describe the dynamics and behaviours of such systems. This process often involves formulating a set of mathematical equations to describe and represent the behaviour of the components, processes, and functions of the system. Usually some efficient and well-established mathematical and computational (inference) methods may be required to

estimate the structure and find optimal parameters of the model from data. For example, in inferring a compound’s mode of action from time course gene expression profiles (Bansal M 2006) used the time-series network identification (TSNI) method to demonstrate how to successfully infer and identify correct regulatory interactions among interrelated genes that are involved in (transcriptional) perturbation experiments (Bansal M. 2005, Bansal M 2006). Such practicable inference of system behaviour may be applicable and useful in predicting network response to external perturbations and identifying genes affected and responsive to drug input. Likewise, Gardner et al. used the network identification by multiple regression (NIR) method (Chua H.N. 2011, Gardner T.S. 2003), an ODE-based algorithm to infer influence interactions (relation of the expression of genes with the expression of other genes) of a gene network, with each interaction implying a regulatory interaction between mRNAs, proteins, mRNAs, metabolites, etc within the cell. Their method too is useful in predicting the network response to drugs and identifying the unknown target genes from experimental data.

Mathematical modelling through time series data analysis has the potential to accelerate new drug design, development, and discovery. We anticipate that such contribution may lead to the development of a new mathematical science. The emergence of contemporary network science (e.g. automated reverse engineering of complex systems using time series data) can contribute immensely to advances in drug design and personalised medicine in cancer research. Through discovery of genetic or other biomarker information from individual data of patient groups, potentially sensitive candidates that may positively respond to treatments (FDA 2012) may be identified and selected. The seamless benefits that diagnostics tools for detecting and mining biomarker information from time series data could deliver would be overwhelming if such detection could be made early and accurately.

An improvement on current data-driven modelling efforts may be required to address some of the most recent challenges in cancer drug development and availability, e.g. 83% of oncologists hit by cancer drug shortages resulting in delayed treatments and associated substantial cost burden imposed by those shortages (drug shortages hit vast majority of oncologists 2013). Complex systems modelling may well reduce the very lengthy process and “out-of-pocket” pre-tax cost of new drugs and apparent concern about how to improve lack of productivity (DiMasi J.A. 2003) may be reduced by supporting the drug development and clinical testing process with improved data-driven modelling or accurate remodelling

of experimental data for informed technical and better decision making.

Mathematical modelling can be used to inform experimental design and enable low-cost hypothesis generation. *In silico* analysis and modelling may be used in system biology studies to address knowledge gaps and eliminate invalid assumptions made about target biological systems (Idowu M.A. 2011*b*, Idowu M.A. 2011*a*, Bown J. 2012).

To adequately deal with system-level understanding and challenges of complex systems (Kreeger P.K. 2010, Bown J. 2012), it might be necessary to develop a modelling approach that both provides new methods for studying and dealing with high-level network inference challenges and helps determine an optimal strategy for identifying a target system that has not been well studied. The modelling approach may involve multimodel solutions and integration, i.e. multiple models may be able to describe the system and, as a result of being consistent in the way they describe the system, may need to be integrated into a single modelling framework. Though such integration might introduce new complexities into the modelling challenges, the additional advantages and apparent benefits such as instant construction of models, faster identification and estimation of systems, and the synergistic effects of the combined power of such integration might create exciting new opportunities for a wider range of solution, e.g. methods that promote instant *in silico* network inference using simulated data may give new hints on the possible of *in silico* modelling and simulation of real complex systems.

As molecular profiling methods are being used to monitor cellular responses to perturbation due to disease or treatments, it is important to apply appropriate data analysis method (Bailey W.J. 2004). We recommend a time series data analysis method that is appropriate for dynamic modelling and reverse engineering of complex systems. The purpose of reverse engineering is to identify (sub)systems and underlying network of interactions from experimental data to promote deeper understanding where little or no knowledge about the processes or underlying principles behind the original system is known. Quantitative time series data (profile) of such complex phenomena may be modelled through network inference. Such target systems are often described as interaction networks of interconnected and interrelated components (nodes, i.e. measurables) and the interactions between any pair of nodes represented by the weighted edges in the networks. Hence the demand for immediate modelling results may be tackled by fast inference algorithms that can provide instant inference of network of interactions from data. We advocate a case for fast network

inference method that may produce instant system identification and parameter estimation results in a matter of seconds to adequately meet the model development challenges often associated with drug design and personalised medicine - effective individual treatment may require personalised dynamic modelling of individual time series data.

To permit real insights into biological time series data, practicable system identification and parameter estimation methods that can infer meaningful results in seconds or minutes are required. Such methods that can translate times series data of complex systems into a representation of the underlying network structures and dynamics of the system and immediately construct network models that allow new predictions to be made are needed. This thesis identifies and validates a novel modelling technique for inferring and extraction useful information from data. Novel matrix-based inference methods are developed and first applied to known simulated test environments to identify fundamental algorithms that are frequently useful in supporting automated identification of systems strictly from their time series data. These new inference algorithms may be considered as fast, inexpensive strategies for solving system identification and parameter estimation problems in mathematical modelling. Though the techniques introduced in this thesis are purely based on time series data, the methods used are flexible enough to incorporate experts' knowledge. The inference methods are generalisable, i.e. the algorithms are not limited by the nature of time series data considered but rather are able to deal with time series data in finance, econometrics, weather forecasting, control engineering etc, just to mention a few.

This thesis presents a new theoretical modelling and visualisation framework able to deal with a wide range of time series data sets relating to complex systems, e.g. biological pathways. The integrated framework developed is able to infer and depict the interaction network implicit in time series data sets and the evolution of that network dynamics in response to treatments. Fundamental research questions that relate to how some key pathways may be regulated in breast cancer are considered also. Here in this thesis, a demonstration of how to model time series data of important biological pathways (e.g. DNA damage response pathway) to automatically and dynamically (i.e. mimic and reflect changes in data in order to) construct predictive models that are consistent (i.e. able to reproduce exactly those same experimental data) is presented. We refers to this process as *dynamic modelling*. We seek to develop new inference methods that can accurately predict both the structure and dynamics of any target system *purely* by analysing the experimental time series data

of that system theoretically and mathematically.

The thesis considers four hypotheses:

1. there exists an integrated modelling framework able to give exact representation of time series data and such techniques are sufficient to produce meaningful solutions to system identification and network inference problems;
2. the framework identified in 1 is *robust* and applicable to a wide range of data (including both surplus and extremely limited data, e.g. data with only 3 time points)
3. the framework identified in 1 may inform experimental design and interpretation in biological systems;
4. the framework can produce an instantaneous result that indicates changes in cell signalling responses to drug action and specifically indicates sensitive and resistant signalling dynamics.

A confirmation (or refutation) of these hypotheses is evidenced in the following chapters.

Chapter 2 deals with a timeline historical account of external factors that may induce cancer formation, a brief introduction to cancer biology, cancer diagnosis, and treatments before introducing cancer systems biology and basic modelling approaches. Background information on the essentials of cancer biology and cancer systems biology are provided.

Chapter 3 considers system representation and a review of several system identification strategies used in the past and parameter estimation methods that have been proposed within the last few decades. The central theme of this chapter focuses on data-driven modelling of time series data based on ordinary differential equations (ODE), particularly biochemical system theory based ODE representation, and the advantages and main challenges of BST and BST based inference methods for modelling dynamic time series data.

Chapter 4 describes the integrative framework developed: analytical methods for modelling time series data; an ODE-based Jacobian method of inference; new matrix decomposition and construction methods; and a power-law based half-system approach for articulating complex systems dynamics. To demonstrate the effectiveness of the Jacobian based method, artificial data are used in the assessment of those inference methods. In this section, hundreds (i.e. 700) of simulated time series data are analysed and tested for the assessment of the fundamental methods that would be used in the actual biological experiments. This chapter addresses hypotheses 1 and 2 and produce scientific and theoretical evidences for their confirmation or refutation.

As an addition to chapter 4, heuristically developed analytical methods are discussed in appendix section F.1. Also new complementary methods for inferring data-consistent, self-reconfigurable nonlinear (power-law based) models from time series data are required, developed, and presented in appendix F. These novel methods may be categorised into two broad groups, namely: direct inference and indirect (recast) methods. The direct method involves applying direct means to infer a jacobian or power-law based model from time series data. The indirect method, however, uses a new system identification method to first infer a jacobian model as instant and temporal solution to the inverse problem before recasting the inferred jacobian model to corresponding power-law model using our newly developed recast technique. The recast method, in addition to normal behaviour, also provides a novel analytical technique for integrating power-law and jacobian models together. This new approach may be used to extend our modelling strategy from matrix-based network inference to model interoperability and multiple model transformation in terms of finding multiple distinct models (solutions) to inverse problems. The structure of this thesis is illustrated in Figure 1-1.

Chapter 5 describes acquisition and analysis of real time series data of the DNA-damage response pathway. Instant analyses of time series data sets are performed to generate heatmap representations of the data and the application of the modelling method to the time series data supplied. This chapter deals on reverse engineering of DNA damage response pathway and application of dynamic modelling in revealing and understanding DNA damage dependent dual consequences of ATM kinase inhibition on cell survival. Hypothesis 3 is addressed in this chapter.

Chapter 6 describes acquisition and analysis of real time series data of PI3K-AKT signalling pathways. Analysis of the supplied time series data sets is the form of heatmap representation of the actual data. The chapter concludes with an application of the modelling methods to generate new results. The results of modelling the PI3K-AKT signalling pathways are then interpreted and discussed. This chapter will confirm hypothesis 4.

Chapter 7 concludes with a discussion of the overall modelling approach and draws out a set of conclusions and recommendations for future work.

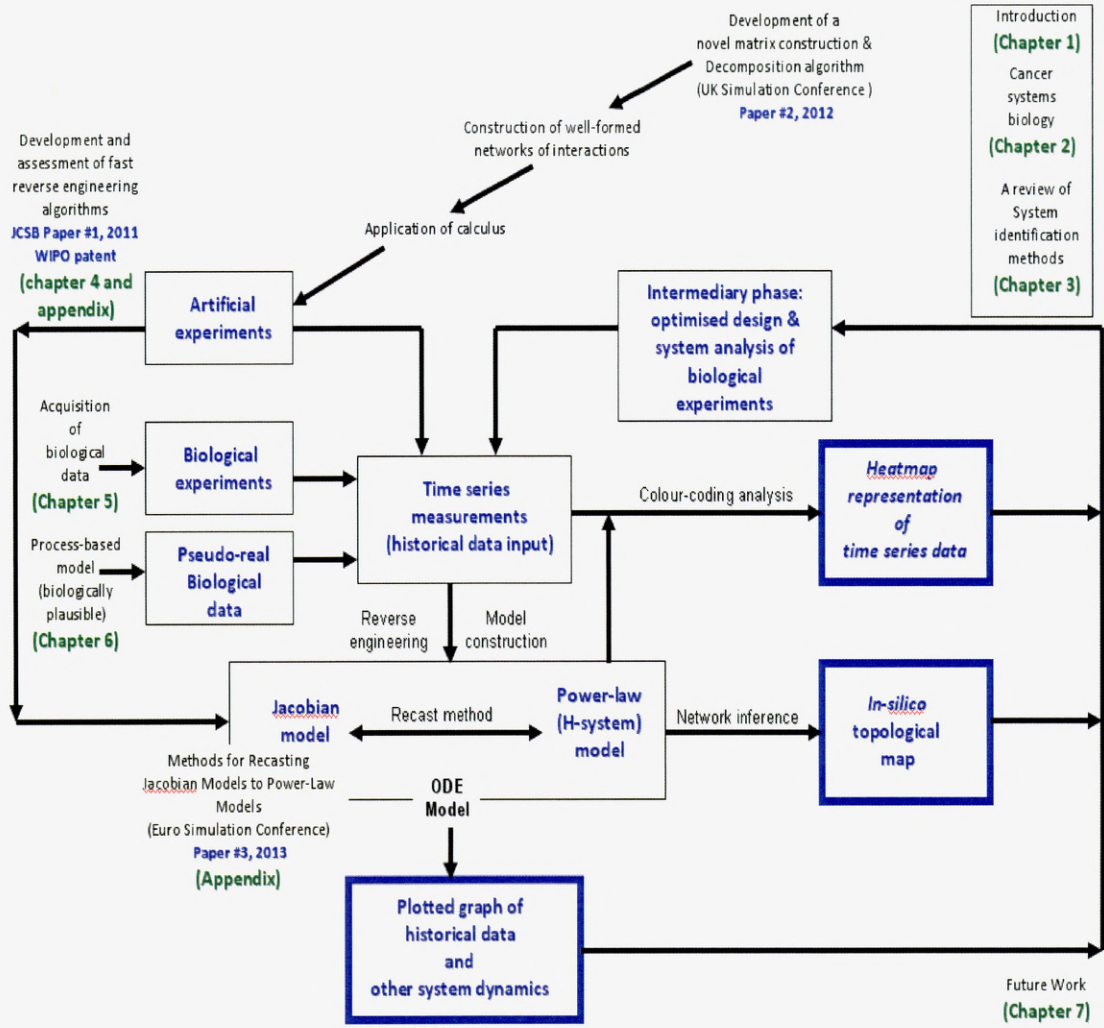


Figure 1-1: Thesis structure.

1.1 Motivation

The need to develop a mathematical and theoretical framework that supports (fast) multimodel integration and automated construction of dynamic (*deterministic, reconfigurable, self-organising, unsupervised, and automatically constructible*) (Idowu M.A. 2011b) models from experimental time series data has resulted in a novel reverse engineering strategy that often guarantees the production of highly convenient, sophisticated and simple models capable of supporting fast data analysis and utilisation for systems identification and analysis purposes. In practice, this involves an automated inference or extraction of unsupervised

predictive *models of time series data* that are useful for making accurate predictions about *other unknown time points and structure of a complex system*.

Chapter 2

Cancer biology and mathematical modelling

2.1 Causes of cancer

Cancer, as a disease of uncontrolled growth and unexpected cell division, is one of the major causes of death worldwide. In UK alone there are over a quarter of a million (around 309,500) new cases of cancer diagnosed each year, and breast cancer is by far the most common cancer in women accounting for almost a third (about 31%) of all female cases in the UK. Just in 2009 alone, there were more than 156,000 cancer deaths in the UK, and over one in four (28%) of all deaths in the UK were due to cancer (UK 2012). Unlike their normal counterpart, cancer cells tend to follow abnormal rules of cell growth and division due to a number of reasons including complex genetic changes, faults and diversity, and identifying and understanding such underlying causes of and treatments for cancer is important.

As a disease involving dynamic changes in the genome, there are compelling evidences to suggest that cancer formation in humans may be a result of evolutionary and progressive changes and transformation in normal cells which eventually may result in defective genetic alterations (Hanahan D. 2000). It may be that almost all human cancers acquire abnormal capabilities that make them almost insusceptible to destruction and insensitive to anticancer drugs. Hanahan and Weinberg propose six fundamental capabilities that often manifest in human cancers, each a feature acquired as a result of defects in regulatory circuits that

govern normal cell behaviour: self-sufficiency in growth signals, insensitivity to antigrowth signals, evading apoptosis, limitless replication potential, sustained angiogenesis, and tissue invasion and metastasis (Hanahan D. 2000).

In the past it was suggested cancer resulted from prolonged conditions induced by some of the above-mentioned factors and other extrinsic (e.g. occupational or environmental) factors such as exposure to soot (scrotal cancer), exposure to X-rays¹(skin cancer), radioactivity, radioactive gas product², smoking and asbestos³, dibenzanthracene⁴(skin cancer), exposure to the dangerous chemicals aromatic amines⁵, exposure beta-naphthylamine⁶(bladder cancer), excessive radiation(leukemia), excessive alcohol intake (breast cancer), excessive exposure to pesticide(increased risk of brain tumours) (D.M 2006). However, there is a general agreement among biological researchers that the following intrinsic factors do contribute to genome instability: presence of oncogenes⁷; defective chromosomal changes⁸ or scrambled chromosome makeup or chromosomal translocation (gene amplification) or deletion, abnormalities within a cell; mutation in genes; gain-of-function mutation⁹; loss-of-function mutation¹⁰; mutated apoptotic proteins; low-levels of or mutated tumour suppressor proteins; overexpression of cyclins; inactivation of tumour suppressor gene; and loss of expression of CDK inhibitors.

In many cases, this information about the mechanistic basis of a given pathology is vital for predicting cellular responses to drug treatments (Kreeger P.K. 2010). Identifying and mechanistic understanding of dysregulated pathways in cancer might significantly lead to optimal experimental design and informed modelling strategies for predicting better treatment outcomes than otherwise.

¹X-rays were found to be both mutagenic and carcinogenic, i.e. able to damage DNA by directly removing electrons and, thus, ionize molecules

²mines of the Ores mountains inhaled by miners

³it was found mesothelioma could be caused by even a low-dose exposure to airborne asbestos leading to the discoveries of two human cancer viruses: Hepatitis B virus; and Epstein-Barr virus

⁴a chemical carcinogen in coal tar applied to the skin of rabbit ears

⁵aromatic amines produced in factories

⁶beta-naphthylamine is carcinogenic but was once present in food

⁷oncogenes, e.g. ras and BRCA-1 and BRCA-2; BRCA-1 and BRCA-2 account for 90 percent of all hereditary breast cancers (and up to 5-10 percent of all breast cancers in general population.

⁸chromosomal translocation can cause chronic myelogenous leukemia (CML)

⁹involves mutation in genes encoding signalling molecule of growth factors, signaling receptors or intracellular receptors, intracellular transducers, and transcription factors, e.g. overproduction of positive regulators such as cyclinD

¹⁰involves mutation in genes encoding cell-cycle control proteins and DNA-repair proteins e.g. loss of (growth-inhibiting) negative regulators such as Rb, p16

2.2 Background on the essentials of cancer biology

Biological pathways, e.g. signal transduction, gene regulation and metabolic pathways, play important roles in biological systems (Kreeger P.K. 2010). These pathways encompass series of important actions and events that produce certain changes and responses in cells, e.g. activation and inactivation of genes, DNA repair in response to DNA-damage in cells, phosphorylation and dephosphorylation of proteins, etc most of which are geared towards appropriating the necessary responses to intracellular or environmental stimuli, right actions in defence, or stimulating new actions in recovery from abnormal negative influences. Contemporary cancer biology assumes that some commonly mutated pathways in breast cancer involve growth-stimulating, growth-inhibitory and DNA-damage response pathways. These involve the human epidermal growth factor (HER) receptor family, Phosphoinositide 3-kinase (PI3K) / Protein Kinase B (Akt or PKB) and Mitogen-activated protein kinase (MAPK) intracellular signalling pathways, Ataxia telangiectasia mutated (ATM)/ataxia telangiectasia and Rad3-related protein (ATR) pathways, or cell cycle control system. In this thesis, fundamental research questions that relate to how these pathways may be regulated or dysregulated in breast cancer are considered.

Just as cancer is associated with heterogenous pathology with respect to tissue and cell type and origin, cancer systems biology is subjective to systematic research in which experimentation and generation of hypotheses are combined. Constant improvement in the understanding of the mechanistic nature of the disease is necessary. A contemporary viewpoint is that the disease involves dysregulation of multiple pathways (Kreeger P.K. 2010). Such views suggest that different malfunctioning components may be involved in a given case and in parallel at the same time (Kreeger P.K. 2010). For this reason, most treatment efforts which were once based on single cause notions ended up producing unsatisfactory outcomes.

Most of those treatments that were based on a single mutated oncogene often either ignored the after-effects or related consequences of those mutation or presumed that the multivariate nature of the molecular level changes involved could still be unwound by the treatment mechanisms suggested. Today, cancer is generally viewed as a highly heterogenous pathology whose molecular network activities are constantly bombarded by alterations and diverse genetic mutations (Kreeger P.K. 2010). An emerging and wisely accepted per-

spective is the necessity to move towards a system-level approach to improve current experimental understanding of the multiple pathways involved.

Focus and research studies may be directed at the level of dynamic protein operations, such as phosphorylation, because this is the level at which most environmental and genomic influences are convoluted (Kreeger P.K. 2010). The nature of the biological pathways being studied depends on the sort of questions or problems being addressed. For example, signal transduction pathways deal with the transmission of signals, gene regulatory networks involve the regulation of gene expression, etc.

In normal cells there is a balance between growth-stimulating and growth-inhibiting signalling pathways. Most cancer cells are developed when normal cellular control circuitry breaks down or fails to function properly. Growth factors eventually signal the cell cycle control system by stimulating DNA synthesis and division. The control of cell growth, cell division, and triggering of cell death are some of the major challenges in cancer drugs treatment. Each of these mechanism is an important subject area that must be studied. The following sections provide background information on the essentials of cancer biology.

2.2.1 Growth factors and the HER family of receptors

A growth factor is a substance that stimulates cellular growth, proliferation and cellular differentiation. There are special proteins, called transmembrane proteins (TP), that are capable of moving from one side of a membrane through to the other side. The human epidermal growth factor receptor (HER) is a family of transmembrane receptors that are involved with the regulation of cancer cell growth and survival. The activation and dimerization of HER family receptors results in the activation of target genes within the nucleus. These genes determine biological responses, such as proliferation or differentiation. They comprise four transmembrane proteins, each with different properties but all involved in the regulation of cell proliferation (C 2003). These receptors are responsible for mediating normal cell growth and differentiation (Alan 1999). Abnormal activities of these receptors have been found to lead to the development of a number of human cancers (E.K 2003). This is the reason why today some anticancer agents target HER receptors.

The four known members of the HER family are: HER1(ErbB1 or EGFR - epidermal growth factor receptor); HER2 (or ErbB2 or c-neu); HER3 (or ErbB3); and HER4 (or

ErbB4). These receptors are typically found on the cell surface of normal tissues of epithelial, mesenchymal, and neuronal origin (Olayioye M.A. 2000). They are structurally similar, but have distinct characteristics that dictate their signalling specificity. They share a structural configuration comprised of an extracellular domain; a single hydrophobic transmembrane domain; and a highly-conserved tyrosine kinase domain.

2.2.2 HER receptor ligand-binding

A ligand is a substance in the form of a molecule or molecular group (e.g. drug, hormone or antibody) that can bind to a target protein (receptor), alter its conformation (state or shape) to trigger a biological signal. HER receptors normally exist as inactive monomers. Activation of the receptor occurs on ligand-binding, and this triggers a cascade of events that leads to receptor dimerization, and ultimately mediates biologic processes such as cell growth and differentiation. Apart from HER2 the other members of the HER family require the binding of an extracellular ligand for activation. Different ligands bind different receptors (C 2003, Olayioye M.A. 2000) and their binding results in different effects (Jones J.T. 1999, W.J 2001). With the exception of Epidermal growth factor (EGF), which is found in many body fluids, the availability of these ligands is one way in which HER receptor activity is controlled.

2.2.3 Conformational change

Ligand-binding leads to a conformational change in the receptor. The receptor changes from a closed state to an open state. In the open state, a region of the receptor known as the dimerization domain is exposed. This allows the receptor to dimerise with another receptor in an open state, and initiate signal transduction (Dawson J.P. 2005) HER2 has no known ligand and its structural conformation is always open, mimicking a ligand-bound state. This allows HER2 to automatically dimerize with other HER receptors, making it the preferred pairing/dimerisation partner for the ligand-activated HER family members HER1, HER3, and HER4 (C 2003).

2.2.4 Receptor dimerization (pairing)

Receptor dimerisation is an essential requirement for HER function and for the signaling activity of all HER receptors. The dimerisation process can occur between 2 different receptors from the HER family (heterodimerisation, e.g., HER1 and HER3) or between 2 of the same receptors (homodimerisation, e.g., HER1 and HER1) (Olayioye M.A. 2000). Stimulation by a specific ligand confers a specific dimerisation profile that is tissue specific or tumour specific (Olayioye M.A. 2000). Dimerisation results in activation of the kinase domain, transphosphorylation, and the induction of intracellular signaling cascades (C 2003). The HER signaling network is highly complex with many possible dimeric receptor combinations, multiple associated ligands, and numerous intracellular pathways. Signaling diversity depends not only on the presence of specific receptors, but also on the characteristics of individual ligands (Olayioye M.A. 2000). Two important signaling pathways activated by the HER family dimers: the PI3K/Akt pathway - promotes tumor cell survival; and the mitogen-activated protein kinase (MAPK) pathway - stimulates proliferation (C 2003). Intracellular signal transduction is initiated by the cytoplasmic domain of the receptor. HER1, HER2, and HER4 all have tyrosine kinase domains, but HER3 has an inactive domain and as a result is unable to directly initiate signal transduction (C 2003). In their inactive, monomeric state, the tyrosine kinase domains of the receptors are not activated. The phosphorylated residues of the cytoplasmic domain act as binding sites for adaptor proteins such as Shc; kinases such as phosphatidylinositol 3-kinase (PI3K); protein tyrosine phosphatases; and guanine nucleotide exchange factors such as Sos (C 2003, Olayioye M.A. 2000). Each receptor possesses a distinct pattern of binding domains, and as a result will form distinct adaptor protein complexes on phosphorylation. This leads to variation in the downstream pathways they activate (Olayioye M.A. 2000).

2.2.5 Cancer and HER receptor dysregulation

Dysregulated HER receptor activity is implicated in a number of tumors, including ovarian, breast, prostate, and lung (C 2003). This dysregulation may be caused by receptor mutation, or over-expression, or the excessive production of ligands. This can result in activation of downstream signaling pathways, leading to uncontrolled cell proliferation; increased potential for invasion, metastasis, and angiogenesis; and decreased apoptosis

(C 2003, E.K 2003). Therapeutic strategies are being developed which target HER family receptors. These agents are primarily monoclonal antibodies which block ligand binding to the receptors, or small-molecule tyrosine kinase inhibitors which prevent signal transduction via the tyrosine kinase domain of the receptor (C 2003).

2.2.6 Intracellular signalling pathways

The formation of a complex of adaptor proteins results in the activation of downstream signalling pathways (C 2003). The receptors in the HER family are linked to the MAPK pathway (Olayioye M.A. 2000). This is brought about through interaction between adaptor proteins and RAS GDP/GTP-binding proteins. Association of the PI3K adaptor protein with receptor tyrosine kinase domains leads to activation of the AKT pathway. The RAS proteins initiate a cascade of phosphorylation events in associated signalling molecules, leading to the activation of the MAP kinases. MAP kinases transfer the signal through the cytoplasm to the nucleus. Intracellular signalling proteins bind to phosphotyrosines on activated RTKs to form a signalling complex, sending multiple signals through multiple pathways, including SRC, STAT, PKC, PLC γ 1/PKC, PI3K/AKT and MAPK pathways (Olayioye M.A. 2000, Alberts Bruce 2009). For the purpose of streamlined focus and data available for this research, we shall restrict our discussion to only the MAPK and PI3K/AKT pathways. The MAP kinase pathway includes the Mitogen-activated protein (MAP) kinases that lie in protein kinase cascades. These kinases comprise a family of protein-serine/threonine kinases, which participate in signal transduction pathways that control intracellular events including acute responses to hormones and major developmental changes. The RTKs may also use an alternative relay mechanism that is responsible for promoting cell survival and growth through the enzyme phosphoinositide 3-kinase (PI3K). PI3K basically phosphorylates inositol phospho lipids (and not proteins). Unlike their protein counterpart, lipids are generally synthesised, modified and broken down by enzymes (Alberts Bruce 2009).

Cell division and regulation

In normal cells there is a carefully regulated balance between growth-stimulating and growth-inhibiting signaling pathways. Growth factors signal the cell cycle control system

by stimulating DNA synthesis and division. The binding of these growth factors to specific receptors on the plasma membrane is usually necessary for cell division (CellLectures 2010). Cyclins and cyclin dependent kinases (CDKs) play a major role in regulating the cell cycle. Signals affecting critical checkpoints determine whether the cell will go through a complete cycle and divide (CellLectures 2010).

The cell cycle control system is a complex and highly regulated system, characterised by temporarily ordered events of oscillations, checkpoints, positive and negative feedback loops. Its control of transitions can be summarised as follows: G0 phase: resting phase; G1 phase: committed to high rate biosynthetic activities and progression through the cell cycle; S phase: DNA synthesis (chromosomes are duplicated); G2 phase: Significant protein synthesis; and M phase: Mitosis (nuclear division - cell divides into 2 daughter cells). The central players of the cell cycle control system are the Cyclin dependent kinases (CDKs), which govern the initiation, progression, and completion of cell cycle events. The control of transition between cell cycle phases is dependent on the quantity of Cyclins, CDKs, and CDK inhibitors. Understanding how these protein kinases may be regulating the cell cycle is important.

2.2.7 Diagnosis and treatment of breast cancer

Breast cancer is currently the most common cancer in the UK (Breathrough 2010). This section aims to give information on breast cancer diagnosis, drug treatments and the potential of computational and mathematical modelling in cancer systems biology aimed at pharmaceutical development.

Treatments given to women with breast cancer include one or more of radiotherapy, chemotherapy, hormone therapy, and targeted (biological) therapy depending on the particular circumstances of the individual. The breast cancer treatment processes can be summarised in three major steps: referral, diagnosis and treatment.

Step 1: Referral

While patients are advised to see their GP immediately if they observe symptoms such as lump, nipple distortion or skin changes (ulceration). Nine out of ten breast lumps are not cancer. The GP will then refer them to a specialist if necessary. Over 80 percent of all

breast cancer cases in the UK are in women over the age of 50. It is important to remember that breast cancer is rare in women under the age of 40 (Breathrough 2010). This age effect is observed since most cancers are formed from a number of abnormalities occurring concurrently. Therefore, cancer may be viewed as a disease formed over a long period of time for multiple abnormalities to aggregate.

Step 2: Diagnosis (Triple Assessment)

The triple assessment (testing for breast cancer) is based on three basic examinations, namely: clinical (physical) examination; breast imaging (mammogram or ultrasound); and core biopsy and/or fine needle aspiration (FNA). Biopsy may be used if the symptom being investigated is a lump. Clinical examination involves a physical examination performed by a doctor or specialist nurse. Breast imaging is performed by using either mammograms (special X-rays that use a very low dose of radiation) or ultrasound scans of the breasts. Core biopsy is performed by taking samples of cells from the lump using a needle. A fine needle aspiration (FNA) - thin needle used to take samples of cells from the breast lump area - may be given before or instead of a core biopsy. The sample is then taken to the laboratory, where it is studied by a pathologist. The results of the tests will determine whether cancerous cells are present.

It is important to know certain characteristics of the breast cancer to determine the best treatment option for the particular patient. Staging and grading are processes (or methods) used to define the exhibited properties of a cancer. The TNM (tumour, nodes, metastases) system of staging describes the size of a tumour, the number of lymph nodes affected, and whether and how far the cancer has spread. Using a scale between 0 and 4, a high number indicates that the tumour is large and has spread beyond the breast and to the lymph nodes or beyond. Tumours are graded between 1 and 3 depending on pathological features, including mitotic rate, tubule formation (ie how similar the cancer is to normal), and the morphology of cancer nuclei. A higher number is associated with a poorer prognosis.

Step 3: Treatments for Breast Cancer

Treatments available to breast cancer patients may be classified as local or systemic depending on staging and grading outcome. Local treatments aim to remove cancer from local sites through surgery and radiotherapy. Radiotherapy involves using radiation to de-

stroy cancer cells. Systemic treatments involve the use of drugs that aim to specifically target and kill cancer cells which may have spread, e.g. chemotherapy, hormone therapy and targeted therapy. Chemotherapy involves using anti-cancer (cytotoxic) drugs to destroy cancer cells. Hormone therapy blocks the production or action of hormones that are considered favourable to cancer. It also aims at reducing the ability of the cancer cells to respond to such hormones. Targeted therapy involves using therapeutic strategies such as HER-targeted monoclonal antibodies, e.g. Trastuzumab (Herceptin), for treating HER2-positive breast cancers. Functional loss of PTEN may be associated with acquired resistance to trastuzumab targeting ErbB receptor family.

2.3 Cancer systems biology and computational Modelling

Cancer systems biology is playing an invaluable role in the understanding of cancer biology today. Computational and mathematical modelling approaches and novel methodologies for analysing data are being used to reveal molecular biomarkers in biological systems. Through contemporary cancer systems biology scientists are gaining new insight in and retaining deeper understanding of complex biological phenomena. As mathematical models of these systems and processes aim to describe the different processes involved in a complex system, they seek to capture their dynamics and explain their behaviour using experimental data. Cancer systems biology is becoming more and more evident through the application of theoretical analysis and generation of new hypothesis. The ultimate aim of the modelling is to determine the optimal therapeutic strategy that has the potential to can collectively trigger mass apoptosis in defective cells or discover new biomarkers that may eventually lead to the reversal of drug resistant responses in during treatments.¹¹

Modern biology is concerned with the understanding of the structures of biological systems both at the systemic and molecular levels. In gaining this understanding the key natural phenomena involved in biological growth, evolution and processes must be understood. This has always been a challenging process depending on the level of limitation imposed on the data capture mechanism or method. Notwithstanding, most of the basic functions of

¹¹Summary of effective therapeutic drugs: ER- and PR-positive breast cancer: Tamoxifen, Exemestane (Aromasin), Arimidex (anastrozole) and Femara (letrozole). HER2/neu-positive breast cancer: Trastuzumab (Herceptin). Advanced HER2/neu-positive breast cancer (breast cancer that have progressed after previous Herceptin treatment): Lapatinib/pertuzumab (not standard care)

biological components and mechanisms of their fundamental processes, which are involved at the genetic, cellular, and organic level, together with those favourable and unfavourable conditions that affect or determine their overall responses and behaviours are now being studied and understood at a scale more than ever before. However, the strategic method for studying cancer biological systems requires more than just an hybridization of the best conventional reductionist approach or most effective holistic approach. Whichever approach that is being used must take cognisance of essential fundamental needs such as instant system identification requirements, fast model construction, data consistency, optimal utilisation of limited data, accurate forecasting, and new knowledge discovery (Idowu M.A. 2011b). The following requirements, if adequately collectively met, should be the foundation upon which effective and workable strategic methods for cancer studies are developed.

We recommend that methods that implement system identification and parameter estimation algorithms should run fast (i.e. in a matter of seconds or a minutes) to be of any practical use.

We recommend that the model construction process (using a data-driven modelling approach) should execute fast just as the system identification and parameter estimation methods are expected to run efficiently too. It seems that to guarantee that an automated process completes its model construction task efficiently such a process needs to implement a non iterative technique (e.g. matrix based methods or solutions). For this reason, an objective to ensure that all methods developed in this thesis is matrix-based.

A model that is capable of simulating (without any error) an exact replica of the original experimental (time series) data that was used to construct it may be regarded as being *data consistent*. We assume that in a deterministic systems most predictive models should be data consistent with historical time series data.

Our experimental time series data studies demonstrate that at least three (3) time points are enough to infer a network of interactions (i.e. transformation matrix) from an unknown system. However, the number of time points required to successful infer and identify a system must be \geq the total number of dependent measurables wiithin the system. Where limited data is available, the system identification method should make optimal utilisation of the limited data a priority.

Accurate forecasting can only be guaranteed only if the predictive model is data consistent, though the converse may not be true under limited data availability.

This is perhaps one of the most important elements of all the recommendations; the primary goal of our modelling efforts would be to discover new information about the target systems. Hence it is important to keep a focus on how new information (that are relevant) may be extracted from experimental time series data acquired from a complex system.

We will return to describe how the computational and modelling framework developed in this thesis seeks to meet each of these requirements in the concluding chapter (chapter 7) of this thesis.

Both molecular biology and genomic biology have evolved over the years and the result of their revolutions is this emerging field called systems biology (Ideker T.L. 2006). To some systems biology may be viewed in terms of investigating the behaviour and relationships of all the elements in a particular biological system (Ideker T. 2001). Its goal is to seek to predict the quantitative behaviour of a biological process under realistic perturbation (J 2003). One emerging trend in most recent definitions about systems biology is the holistic approach that must be adopted - seeking to understand and predict the behaviour of biological systems at the system level (Ideker T.L. 2006), (Ideker T. 2001), (J 2003).

A good and intriguing definition of systems biology coined around its operational components is the description “Measurement, Mining, Modelling, and Manipulation”¹². In recent times, more and more multivariate data of biological processes are being captured using advanced high-throughput technologies such as the reverse-phase protein microarrays (RPPA). After data acquisition computational algorithms will be required to mine and generate hypothesis from such data and consequently computational modelling are then used to develop new predictions (Ideker T.L. 2006). These predictions may be useful in informing new experimental design. Ultimately, forms of experimental manipulations or biochemical interventions (Ideker T.L. 2006) must be used to test those predictions produced from computational modelling. From this description, data acquisition supported with effective computational algorithms and good modelling techniques for making predictions help in making contemporary systems biology a powerful scientific subject. With the advent of high-throughput technologies for capturing data from assays, high-throughput analysis of intracellular signalling data may be produced.

Cancer systems biology focuses on understanding the molecular interactions and characteristics of cancer cells. From a computational perspective, the use of computational

¹²definition and illustration taken from the <http://csbi.mit.edu/> webpage

models of cancer consistent with experimental data is required to uncover new insights into cancer mechanisms. Such models are powered by computational and mathematical methods for identifying systems and estimating best parameters that will produce realistic models. These system identification and parameter estimation challenges requires fast and efficient computational methods to be developed. Human cancer systems biology uses predictive models of cancer to describe and understand the behaviour and nature of the disease in humans to influence drug discovery (Butcher E.C. 2004). To help improve decision making in pharmaceutical development, emergent properties of complex human cells are captured and integrated into the relevant drug discovery process, enabling clinical indication selection (Butcher E.C. 2004).

At the cell signalling pathway network and cell-cell interaction scales, systems biology in the pharmaceutical industry may focus on the identification and measurement of molecular components, generate data from high-throughput assays at multiple interactive pathways to address cell responses to physiological stimuli and pharmaceutical agents, develop and use an appropriate cancer models designed to address specific questions at either the pathway or organ level (Butcher E.C. 2004). Focusing on the building blocks of biological systems (genes, proteins, metabolites), data derived from such components may be analysed to identify new targets and generate testable hypotheses, informing experimental design, accelerating drug discovery, and validation of drug efficacy under specific conditions.

Drug approval rates still lie far below the cost of new drug discovery (?), demonstrating that cell biology is expensive research - huge investment into genomics and screening technologies is required. It is believed that computational systems biology might help improve these rates and achieve greater impact on target validation and clinical development decisions.

Computational systems biology requires the integration of experimental and computational research to understand complex biological systems including practical innovations in medicine, drug discovery and engineering. (H 2002a). Its primary goal is to provide a framework for the generation of new hypotheses and accurate prediction based on *in silico* simulation of related disease biology (Ideker T. 2001, Ideker T. 2003). Often faced with the problem of lack of adequate data the adopted system identification strategy encounter great challenges. A lot still remains unknown about organ and system-level responses, even if cell-level responses behaved as expected, and without this knowledge the disease-relevant

biology cannot be integrated into the drug discovery process. To address the problem of inadequate data more costs will have to be incurred in the purchase of high-throughput technologies furthering the rate at which investment cost exceeds that of benefits. Difficult balance must be struck between outstanding modelling outcomes and data limitation.

Still other factors such as tumour heterogeneity have the potential to make cancer drug discovery process an even more challenging endeavour (Alexander Kamb & Lengauer 2007). The “war on cancer” is far from being won as organ level understanding is multi-scale, requiring knowledge of operations that occur dynamically at gene, pathway and cell levels. Since cells may be represented individually, questions of spatio-temporal heterogeneity such as how to predict tumour behaviour and response to intervention in spatially distributed mediums must be addressed (Salvatore Pece & Fiore 2010). In addition to this, multi-scale models are extremely difficult to construct and integrate due to the levels of uncertainties and change in the underpinning knowledge base, model purpose and scope (Bown J. 2012).

For model complexity to be interpretable, “models need to be as simple as possible but no simpler”¹³, i.e. adequate care should be taken to ensure that where model kinetics and parameters may be aggregated into smaller number of representative parameters or constants, they should be aggregated into understandable components. Another way of promoting model simplicity is by streamlining the model’s scope, covering and focusing on only the questions to be addressed by the model (Idowu M.A. 2011a, Bown J. 2012).

Another great challenge of recent systems biology and contemporary medicine is related to personalised medicine or healthcare solutions, which involves an identification and application of an individual (patient’s) biofeatures to drug therapy discovery, efficacy modelling and healthcare management (J.K 2006). This would require biomarker diagnostics, data analysis, information extraction, visualisation of association network of system kinetics (Idowu M.A. 2011b). In this approach health services and coherently tailored and prescribed therapeutics may be made available to individual patients or defined sub-populations applying recent knowledge obtained from pharmacogenomics and clinical practice (A.M 2007).

Therefore, it is important to understand that the overall goal of modern systems biology is to understand physiology and disease from the level of molecular pathways, regulatory

¹³One of Albert Einstein’s quotes: “It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.” - “Everything should be made as simple as possible, but no simpler.”

networks, cells, tissues, organs and ultimately the whole organism (Butcher E.C. 2004) requires pulling together broad resources and scientists from many disciplines (A.R. 2008).

2.4 Process-based modelling

When investigating how multiple interrelated components of a complex systems interact mathematical models can be used. Conventionally, there are two basic classes of models, depending on the modelling approach or how the models are created and used: process-based models and data-driven models. A process-based modelling approach is one strategy often used to capture and describe the important mechanistic details of the underlying processes and fundamental behaviour of a complex system producing process-based models during the process. Such models are then used to investigate emergent properties that are often impossible to infer intuitively (A.R. 2008).

Process-based modelling often involves very difficult challenges due to the nature of the complications that may arise during the integration of all key processes, especially on multiple scales - the fundamental principles governing the various interactions must be understood, captured and well formulated, each of these being a challenging tasks on its own. All the various complexities involved must be considered e.g. gene to gene interactions in gene networks, signals transduction in mutiple signalling pathways all the way up from the molecular level to higher biological scales at the population level (A.R. 2008). Predictive individual-based models of cancer cells e.g. models of signalling (or gene) networks may be used to predict key biomarker functions. The potential to introduce other more complicated modelling questions or challenges such as how to handle genotype-to-phenotype mapping to predict the behaviour of a cancer cell introduces new insights into the complexity involved in the challenge. One major problem that is yet to be tackled in process-based modelling is the challenge of bringing multiple processes that cut across multiple scales in quantitative terms without combining facts with erroneous assumptions that have the potential to affect modelling outcome in a drastic way (A.R. 2008).

The construction of “highly scalable models of biologically plausible cells arranged in biologically plausible structures that model cell behaviour (lifecycle), interactions (biomechanics) and response to therapeutic interventions (cellular signalling)” has been suggested (Bown et al. (Bown J. 2012)) in response to addressing the fundamental goal of

systems biology - that is, the requirement to address and promote biology (or pathology) understanding at the system level also (Ideker T.L. 2006), (Ideker T. 2001), (J 2003). However, major challenges such as dealing with complexity in multi-scale modelling, uncertainty and change in the underpinning knowledge base, and uncertainty in the model purpose and scope (Bown J. 2012) should be addressed. In addition to these, it has to be decided which type of modelling (functional) form is most appropriate for the particular system being investigated.

It is common to use the Michaelis-Menten rate laws approach (Briggs G.E. 1925) in process-based modelling. Generally, the key biological processes are first investigated and then formulated with appropriate mathematical functions to describe and incorporate then with kinetic equations and parameters. For example, cell processes may be described by modelling and approximating the relevant and important underlying intracellular processes. Though such process-based models are often difficult to create and complicated to estimate parameters, they are able to capture the necessary information about the various processes that should be integrated. Process-based models must make the right assumptions about all the key associations, dissociations, internal processes and external influences within a given system. For example, to identify molecular targets, computational systems biology may use computational modelling approaches that integrate cell structure and dynamics (H 2002*b*). Initial knowledge about cell activity, processes or responses may be extracted from the biological experts and literature. The topologies of the networks (and their components) ought to be known and specified. Such topologies and all the related associations amongst the systems component are intimated in sets of ordinary differential equations (ODEs). The ODE model is then used to describe the rates of change in the biological components, which will be used in simulating the dynamics of all the individual components and the whole. Such models have the potential to offer real insights into the missing knowledge between biological mechanisms and signalling responses (Bown J. 2012), (B 2006), thereby assisting in providing a platform for promoting understanding of how these signalling networks work.

2.4.1 Major challenges of process-based modelling

Since cancer is characterised by abnormal activities of (multiple) pathways, it may serve well to characterise signal transduction in a multi-pathway network, where cell processes are con-

trols by those signals (B 2006). Recent target-based cancer drug development approaches are now focusing on the aberration of the interconnecting network of cellular signalling pathways involving ligands, transmembrane receptors, intracellular signalling protein kinases, and transcription factors (Adjei A.A. 2005). In cellular signalling models, the networks that connect these multi-pathways are revealed to have highly complex topologies with feedback loops (Papin J.A. 2005), crosstalk, possibly alternate interconnectivities, often enabling cells to be robust to perturbation (Bown J. 2012). Hence it is almost impossible to keep track of all these activities reducing model identifiability and creating a huge gap in knowledge about model construction and system processes formulation. The potential for making wrong assumptions and outdated assessment of system states is great, not mentioning the difficulty in dealing with data from different scales.

Due to the level of complexity involved in formulating the essential cellular functions in signal transduction systems, it is important to seek alternative methods that can be used to extract useful predictions from experimental data for complex cellular signalling networks (Brown K.S. 2004). Both the choice of model to be used and the modelling method adopted should depend on the nature of the biological questions to be addressed. Important issues such as data limitations, multi-scale integration challenges, false assumptions must be considered before choosing the appropriate modelling formalism to be used (Morris M.K. 2010). These are some of the issues addressed in this thesis.

2.5 Data-driven modelling

The advent of high-throughput technologies and equipments has contributed immensely to the acquisition and development of large-scale quantitative studies of signal-transduction networks. Such data are hard to understand completely by inspection and intuition (Janes K.A. 2006). In analysing such large data sets, data-driven modelling approaches may require developing computational algorithms and methods for analysing experimental data. Data-driven modelling is an approach in which system identification and model construction and calibration depend on a given experimental data and consistency with that given data. With careful optimisation, adequate and efficient analysis of such acquired measurements the constructed data-driven models may help extra useful information from experimental through reverse engineering. Modelling frameworks that incorporate data-driven models

and methods are fast becoming important platforms for systems-level research in signalling networks (Janes K.A. 2006). Data-driven modelling is described fully in chapter 3.

This integration of biological data into mathematical models often involves time series measurements following some stimulus e.g. gene expression, protein concentration, or metabolite concentrations (Voit 2008). Information extraction or data mining of time series data sets is aided through the development of fast and efficient computational techniques for reverse engineering experimental data. During the process, data-consistent models are constructed using network inference algorithms that may or may not make use of *a priori* assumptions about network interconnectivity. Most of the time, these computational algorithms struggle in ensuring that the constructed model is consistent with the given experimental data (Voit 2008). Some optimisation work might be required to ensure that the network inference algorithms produce fitness results that converge rapidly fast, reach true global optimum, do not overfit noisy or partially missing data (Voit 2008). This effective management of experimental data is essential in pharmacokinetics. Computational systems biology will continue to support drug design i.e. experimental design and clinical decisions will continue to be informed by the results obtained from data-driving network inference.

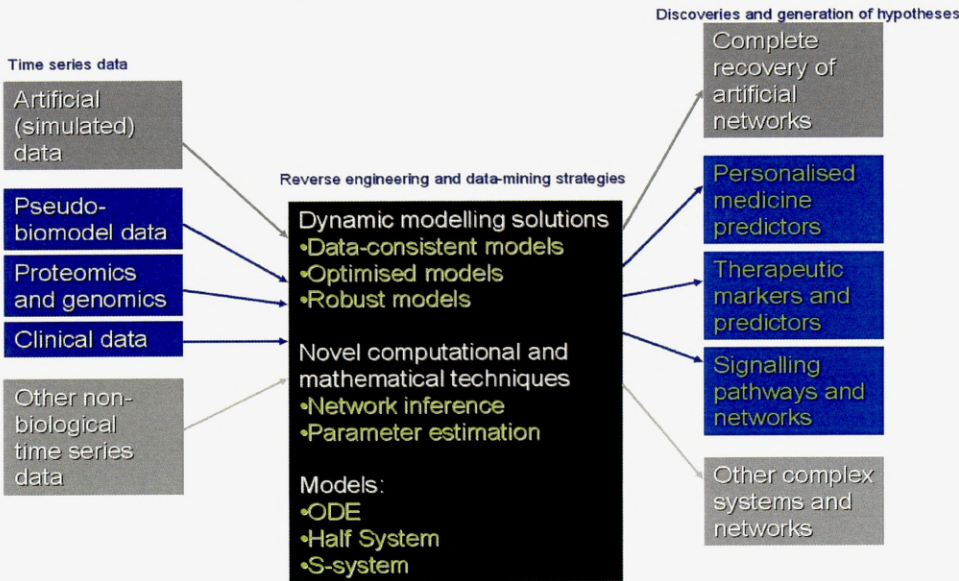


Figure 2-1: A proposed robust and inexpensive matrix-based reverse engineering framework that is able to optimally utilise limited time series data.

Experimental protein interaction data by high-throughput techniques usually have many false positives; specifically, proteins that do not interact in reality are observed to interact

in the experiments (Chen L. 2009). As a result, many of these presumptions are made in mathematical models, rendering such models inconsistent with new experimental data. On the other hand, because of the incompleteness of an experimental dataset, there may also be false negatives; that is, proteins that interact in reality are not observed to interact in the experiments (Chen L. 2009). Other challenges include the volume of experimental data required for some methods such as Bayesian inference, artificial neural network (ANN), running time of estimating parameters, difficulty in interpreting solutions biologically - no causation; no mechanisms. For this reason, dynamic and more flexible computational methods and network inference algorithms are required.

In cells, pathologies emerge as a result of changes to relatively few pathways in the network of signalling, biochemical and transcription processes. Whilst the cell is robust to perturbation of most of the pathways, correlated perturbation of a comparatively small subset can lead to profound changes in the integrated dynamical behaviour of the cell. Our investigation addresses the question of whether time series data may be used to understand cancer in terms of a dynamic cell network, investigating how the results of time series data analyses may be used to identify biomarkers from tissue samples, employing dynamic models driven by real experimental data focusing on gaining new insight into multiple-drug target interventions.

A data-driven modelling strategy is an approach which is based on the implementation of artificial intelligence and/or useful information extracted from data. It allows the development of inference methods to determine the structure of the interaction network that best approximates the key characteristic features of the system that produces the source data. Hence a constructed model produced from a data-driven modelling strategy seeks to incorporate a simulation intelligence and capability to reproduce (simulate) the original (source) data. We seek to develop and optimise only inference methods that are capable of demonstrating complete recovery of artificial networks of interactions purely from artificial data. A further optimisation of those methods may be required for unsupervised reverse engineering of biological time series data. The figure above (Figure 2-1) illustrates the reverse engineering strategy adopted in this thesis. As illustrated in the figure, in contrast to most traditional process-based modelling approaches, the data-driven modelling approach we have adopted does not require static (fixed) structures of discrete symbols or components to be formalised and does not require continuous human intervention (because it requires

no additional *a priori* assumptions to be specified as inputs) during the development process. The approach we seek to develop is ODE based, completely data-driven (i.e. purely formalised based on the states of the system measured at some given time points and the dynamics of the systems over a time period), and will require a deterministic, continuous model to be formulated analytically and mathematically. The model construction process may be completely automated.

In all case studies, the data-driven strategy used will seek to optimally utilise the given time series data obtained from real experiments or manufactured from an independent process-based model of important pathways (i.e. DNA damage response pathway or PI3K/AKT/MAPK signalling pathways). The algorithm developed and presented in chapter 4 will be used to automatically create the dynamic models that are consistent (i.e. able to simulate or reproduce the exact time data series data input), and ultimately seek to predict both the structure and dynamics of the biological systems (Idowu M.A. 2011*b*, Voit 2002*b*).

Chapter 3

System identification methods

3.1 Data-driven Modelling Approach

In this section we introduce system representation approaches with system identification and parameter estimation challenges. First we introduce the concept of reverse engineering, particularly those based on power-law formalism. The power-law formalism of the biochemical systems theory (BST) is identified as a more effective alternative to the well-known Michaelis-Menten formalism (Briggs G.E. 1925). The BST may be viewed as a mathematical modelling theory for describing complex systems. Stressing the need to reduce model structure restrictions to encourage robust model reconstruction, we review common modelling challenges confronting contemporary modelling and revisit how modelling challenges and issues were addressed in the past. We identify some of the effective data-driven modelling techniques used in the past and, in a chronological order, consider some key issues that have emerged over the years and relate them to why the core method developed in this thesis is important. By discussing some of the modelling developments that have emerged over the period of the last decade, we recount some of the failures and successes of reverse engineering strategies, particularly those based on BST. With a view to working towards fast and immediate reconstruction of dynamic models from time series data, we reaffirm the use of synthetic benchmark data for the support and development of effective, fast and robust inference methods as a useful technique for carrying out risk-free assessment of potential and competing inference methods. Finally, the BST is reintroduced in a much deeper detail by introducing the Half-system, S-system and the generalised mass action (GMA)

systems as some of its forms.

This chapter as a whole focuses on data-driven modelling of time series data based on ODE, particularly BST based ODE representation, and the advantages and challenges of BST and BST based inference methods for modelling dynamic time series data.

3.1.1 ODE based modelling

Ordinary differential equations (ODEs) are commonly used to describe dynamic systems. Whenever time series profiles of constituents of a complex dynamical system become available, such time-evolution dynamics may be described either by a set of ODEs, e.g. jacobian or power-law model. Such time series evolution may be described in mathematical terms to capture system behaviour and states recorded at various time points and intervals. One of the most difficult challenges in modelling biological systems from time series data is the determination of a data-consistent solution to its model reconstruction problem, i.e. inverse problems. Solving an inverse problem often requires developing or applying system identification and parameter estimation strategies to (re)construct a predictive model and calibrate its parameters in such a way that the overall model itself may be workable and consistent with experimental data. Depending on the nature of the systems, the modeller may adopt a power-law model, either as a complementary approach to other existing approaches or as an alternative means to formulate and validate system behaviours or dynamics through modelling. In describing complex biological processes through modelling, one must take into account a consideration of the underlying nonlinear phenomena involved and all the essential relationships among the system components. S-system (Voit 2008, Voit 2000) is an example of an ODE model that may be used to approximate and articulate complex system dynamics in meaningful ways.

3.1.2 Systems representation, identification and parameter estimation

In studying biomolecular networks, computational techniques may be used to analyse and model data to reveal important mechanisms. An understanding of some of the various computational methods that others have used in the past is important. It seems that many modelling strategies, including those developed in recent times, are either based on further

development of older techniques that require some form of optimisation or newer approaches developed specifically to further address key context-sensitive questions. It could be said that many of those methods, in trying to find the best solutions to difficult questions, worked reasonably well in addressing the key modelling challenges that had to be tackled. In fact, a knowledge of some of their foundations is essential for the development of newer techniques. However, it is important to reveal one key factor we have identified that appears to be missing in most modelling approaches. That is the need to develop a mathematical and theoretical framework that supports multimodel integration and automated construction of dynamic models from experimental data. In practice, this might be an automated and immediate inference method or strategy for inferring multiple context-sensitive models from single datasets; a system for recasting a jacobian model to either power-law Half-system or S-system; or any other similar process for promoting cross-platform integration of modelling approaches. An appendix section F.1.1 contains a full description of matrix-based analytical methods for recasting jacobian models to power-law models (Idowu M.A. 2013).

Network systems biology, a vital aspect of systems biology which deals with the understanding of biomolecular networks at gene, protein, cell, tissue, or organism level, tends to rely on information from experimental data. Through application of computational methods and mathematical models the important biological functions of cellular systems and details of their underlying network interactions can be revealed. From a system or network perspective, data obtained from gene regulatory networks, transcription regulatory networks, protein interaction networks, metabolic networks, signal transduction networks, and integration of heterogenous networks may be used in modelling studies, e.g. inferring topological information from data of such a system or network (Chen L. 2009). For example, the formation of a global view of cellular function requires a complete measurement of the expression of all the genes in a cell. This gene expression profiling can be used to study the regulatory relationship between genes. Fast system identification may help provide quick-and-dirty estimates of interaction networks from available data to formulate new hypotheses for further testing e.g. inference of signal transduction pathways and drug targets from data of perturbed experiments (Chen L. 2009). New data generated from such a network model may be used for predicting and analysing system dynamics to aid in the understanding of how the system works.

This inference of relevant information often poses difficult challenges and involves in-

corporating existing methods or developing and integrating new computational tools and methods for modelling experimental data and analysing modelling results to generate *in-silico* topological maps for understanding biological processes at the genome or proteome levels.

Genome level investigation involving molecular network studies on gene regulation and expression in transcriptional regulatory and gene regulatory networks may be conducted, e.g. measuring the products of transcriptional regulation, investigating the interaction and effects of a transcriptional factor in promoting or activating the recruitment of RNA polymerase to specific genes¹. An investigation into the various roles that transcription factors play, how they regulate gene expression, how RNA polymerase function in transcription of genetic information from DNA to RNA etc these are some of the important issues that may have to be addressed through modelling of experimental data. The result of such investigation may turn out to be different from one experiment to another, e.g. interaction between a promoter and RNA polymerase could turn out to be negative, i.e. indicating some repressive effects.

Another important area in which modelling might help is in the area of quick simultaneous detection and analysis of multiple mRNA expression levels during Microarray experiments. Modern-day Microarray techniques and technologies such as the DNA microarray (genome or DNA chip) are now being used to produce large quantities of data. The dynamics and underlying biological processes involved in those data can be inferred. For example, mRNA synthesis and degradation may be better understood by modelling those large amount of data produced during microarray experiments.

Other investigations might involve protein modification, complexes and pathways formation, or other such protein-protein interactions that may form a protein interaction network at the proteome level (Chen L. 2009).

Here, the interaction between components (nodes) of the transcription regulatory networks, i.e. gene products (mRNA, transcription factors and other protein) and genes, may be represented as the edges of a representative gene regulatory network diagram (Brazhnik P. 2002). The modelling challenges in this context include inferring both the

¹Activators enhance the interaction between RNA polymerase and a particular promoter, encouraging the expression of the gene. Activators do this by increasing the attraction of RNA polymerase for the promoter, through interactions with subunits of the RNA polymerase or indirectly by changing the structure of the DNA

indirect and direct interactions between the different genes and proteins from data in a fast and consistent way, making allowances for gene, protein and metabolite spaces to be integrated into a single network diagram (Brazhnik P. 2002). This last point requires thinking in terms of new modelling concepts and logics that is both flexible and robust in structure. An objective of the modelling task may be to determine how by both internal and external signalling some particular genes are being transcribed.

Essentially the nature of the network to be studied will determine the type of interactions that are involved, e.g. transcription factors to DNA interaction (transcription regulatory networks); gene to gene interaction (gene regulatory network); protein to protein interactions (protein interaction network); enzyme-substrate interactions (metabolic network); molecule to molecule interactions (signalling network). So as here, if the nature of the network concerned is metabolic or part of signal transduction, then network-based studies of interactions, pathways, and subnetworks may be considered as basic components. Often the modeller's aim is finding and applying efficient method to analyse and model either transcriptomic, proteomic, or metabolomic data to discover or reveal some essential biological mechanisms in cellular systems. The overall challenge might be understanding complex system functions from a system viewpoint.

Gaining system level understanding requires overcoming the difficulty and complexity involved in dynamical interaction network of genes, proteins, and biochemical reactions (Chen L. 2009). In this regard, it helps to know about the computational methods in systems representation and identification of biological systems.

On the basis of experimental data generated from biological systems, effective computational methods can help provide deep insights into the mechanisms of cellular systems (Chen L. 2009).

It is common to express or model biological networks in terms of ordinary differential equations (ODEs). However, in situations where the actual architectures of the network model must be derived from data, the types of ODE models for capturing all the important nonlinear functions and processes hidden in experimental data should be flexible, robust and generic enough to enable dynamic modelling in concept with some level of abstraction. Details of how this may be achieved will be discussed in chapter 4. However, it is important to identify the contributions others have made in the area of dynamic modelling using data.

3.2 A review of deterministic modelling approaches

A deterministic modelling method may conceptualise outcomes of causality in a system by seeking to predict those outcomes based on some supportive scientific and theoretical evidences. In formulating an automated deterministic model, the model structure and parameter search space domain may be adapted to recalibration based on known principles, theories, or formalism. The choice of the parameter estimation strategy employed may be directly conditioned to whether or not a right modelling approach is used. A review of existing modelling approaches and system identification methods is presented before explaining the data-driven modelling approach developed in this thesis.

According to Irvine (D.H 1988) the biochemical system theory (BST) may be considered as an efficient modelling framework for analysing nonlinear models due to the availability of efficient recasting and analytical methods built around it. The BST or power-law based framework, which includes Half-systems and S-systems, was originally developed for analysing organisationally complex systems and is ideal for representing growth and development patterns, genetic circuits, immune networks, ecological interactions etc (D.H 1988). They can be used to conveniently represent complex systems (e.g. molecular and cellular networks) and quantifiable elements of such systems using special functions in biophysics and physics (e.g. rate laws for enzyme kinetics, growth laws, probability functions, Cobb-douglas production functions (in economics)). For example, general methods exist within the S-system framework for finding steady-state solutions and performing sensitivity and stability analyses. Hence the BST formalism is naturally a good theory for modelling biological systems. Sorribas and Cascante et al. (Sorribas A. 1994) used a power-law model (solution) to identify a metabolic pathway using dynamic data and steady-state measurements. Hernandez-Bermejo et al. (Hernandez-Bermejo B. 1999) presented a power-law model derived by a least-squares (LS) minimisation criterion as a data-consistent alternative to other traditional derivations. They further extended its definition to include more operating points (Hernandez-Bermejo B. 2000). The foundation of the power-law formalism could be traced back to Michael A. Savageau (Ni Ta-chen 1996) who presented it as an alternative to other traditional formalisms like the Michaelis-Menten formalism (Briggs G.E. 1925) when modelling metabolic pathways of the human red blood cell. Alves and Savageau (Alves R. 2000) introduced the concept of mathematically controlled compar-

ison to differentiate between two promising candidate models (i.e. workable solutions) using robustness and stability measures that determine the sensitivity to parameter fluctuations profiles and deviation from and/or return to steady state after a small perturbation. Voit and Radivoyevitch (Voit E.O. 2000) used the BST modelling framework to model systems with DNA microarray data and enzymatic process information. Voit (Voit 2002a) presented and recommended BST as a modelling framework for processing and analysing large amounts of experimental time series data of genetic and metabolic data. Voit is a leading expert on mathematical modelling using the BST. Tournier (Tournier 2005) modelled a subsystem of a real metabolic pathway with S-system based on the stability analysis of the steady state and relationship between the kinetic parameters of the model. An iterative process is adopted to numerically compute positive equilibria. With proposed conditions for the existence of a unique equilibrium in the phase space firmly established, the S-approximation algorithm used involves symbolic computation of partial derivatives. The use of context information supplied by the biologists is encouraged to derive a piecewise approximation of biological systems. Gonzalez et al. (Orland R. Gonzalez 2006) proposed a parameter estimation algorithm called simulated annealing (SA) to infer S-system models from time series data of biological systems, e.g. signal transduction, gene regulatory and metabolic networks. Polisetty, Voit, et al. (Polisetty P.K. 2006) presented an inference method called branch-and-reduce to tackle global optimization challenges identifying metabolic system parameters using the generalised mass action (GMA) models. This method is suggested to be a better alternative to genetic algorithm (GA), SA, and even most nonlinear regression. Voit (Voit 2005) introduced the smooth bistable S-systems as a possible answer to questions related to multiple stability conditions necessary for capturing and communicating switching phenomena that may be observed in cell cycle control, gene expression, signal transduction or similar. Voit proposed how bistability representation could be achieved by suggesting piecewise power-law approximation using the magnitude of a model parameter to determine and control the internal structure of systems. Chou, Voit, et al. (Chou I-Chun 2006) proposed the alternating regression (AR) method of parameter estimation in biochemical systems models. AR, combined with methods for decoupling ODE, provides fast and effective tool for estimating the parameters of S-system models from time series data. This method works through a decoupling of the ODEs to allow system parameters to be estimated one equation at a time using concentration and slope values of each dependent

variable. A major drawback of AR is its frequent convergence issue. Vilela, Chou, et al. (Vilela M. 2008) proposed a novel parameterisation method for identifying S-system models from time series data without requiring *a priori* topological information to be specified based on eigenvector optimisation of a matrix formed from multiple regression equations of the linearised decoupled S-system. With further extension the method is able to add constraints on metabolites and fluxes, and using synthetic time series data, demonstrates an effective, automated reverse engineering strategy for identifying correct network topology from a collection of other data-consistent models. Rosario and Voit (Rosario R.C.H.d. 2008) investigated and evaluated the performance of a lin-log method for modelling the glycolytic pathway in *Lactococcus lactis* using *in vivo* time-series data. In this approach dealing with several variables that approach low concentration (and ultimately zero) values is challenging. Ko, Voit and Wang (Chih-Lung Ko 2009) introduced a constrained optimisation technique for estimating parameter values of GMA models using both time series measurements and steady-state data. Such constraints, which are based on the flux connectivity information of the system at the steady state, may help produce more accurate representations of model parameters than in unconstrained conditions.

Tominaga and Okamoto (Tominaga D. 1998) developed and applied a GA method to infer S-systems models. The method may require a time consuming effort. Kikuchi et al. (Kikuchi S. 2003) used GA to optimise the parameters of S-system models of small-scale genetic networks based only on time series data of gene expression - a strategy presented to predict both network structure and dynamics using a unified extension of GA and S-systems. This method requires large quantities of data to work. Nyarko and Scitovski (Nyarko E.K. 2004) presented a method for solving identification problems using GA and second-order ODE model. Spieth et al. (Spieth C. 2004) introduced a GA-based memetic inference method for modelling gene regulatory networks based on S-systems in order to avoid cyclic network disallowance in bayesian networks. Kimura et al. (Shuhei Kimura & Konagaya 2005) presented a method of inferring S-system models of genetic networks using a cooperative coevolutionary algorithm and measured time-series data of gene expression obtained from a fairly large network system. The coevolution algorithm breaks a fairly large inference problem into smaller manageable subproblems that can be solved simultaneously. However, the method is based on an iterative process and a form of clustering strategy may be required to analyse large-scale datasets. The method might converge to

a local minimum and cannot no guarantee of actual parameter estimates. Noman and Iba (Nasimul Noman 2005) presented an improved evolutionary method for inferring gene regulatory networks using S-system and differential evolution (DE) based on *in silico* time series data. They used multiple artificial time series data. Later (Nasimul Noman 2006) they proposed an information criteria based fitness evaluation to select model instead of the traditional mean squared error (MSE) fitness test using both small and medium-scale artificial networks during verification. They used S-systems to model a genetic network and successfully identified the network topology and kinetic parameters. with larger networks decoupling of S-system solution may be required. Searson et al. (Searson D.P. 2007) presented S-systems with evolutionary algorithms to infer chemical reaction networks from fed-batch reactor experimental data with limited input of *a priori* knowledge about products and reactants.

Akutsu et al. (Akutsu T. 2000) proposed the linear programming (LP-) based method to infer S-systems based models from time series data using finite differences of derivatives. This method also depends on availability of large quantities of time series data to work. Diaz-Sierra and Fairen (Diaz-Sierra R. 2001) proposed a nonlinear optimisation algorithm to estimate kinetic parameters using time series data and approximate system responses to perturbation from steady-state using multilinear regression. Moles et al. (Carmen G. Moles & Banga 2003) attributed failure in inverse problems to traditional local optimisation techniques and proposed evolution strategies (ES), a form of stochastic algorithm, as deterministic and stochastic global optimization methods. Veflingstad and Voit (Veflingstad S.R. 2004) proposed preprocessing time series data and fitting them preliminarily by multivariate linear regression to improve initial guess quality in priming network inference algorithm. The method used is based on Taylor's theorem and depends on using steady-state data and linearisation techniques. Srividhya et al. (Srividhya J. 2007) presented a global nonlinear modelling technique that generates a complete dictionary of polynomial basis functions based on the law of mass action to construct biochemical pathways from time course data. Tucker and Moulton (Warwick Tucker 2006) presented the method of interval analysis to construct and estimate S-systems based model of a metabolic network. This completely deterministic method allows an exhaustive search of the domain of all parameter values within a finite number of steps circumventing global minimisation problems using a pruning scheme that is based on a boolean function (the cone condition). They used a vector-based

technique to discard unrealistic network topologies from the solution set.

Almeida (J.S 2002) proposed using the artificial neural network (ANN) method, which is a form of artificial intelligence (AI) and machine learning technique, to identify complex relationships from experimental data of biological systems with applications to expression profiles, genomic, and proteomic data aimed at computer-aided medical diagnosis and biological sequence analysis. Almeida and Voit (Almeida J.S. 2003) presented a data smoothing method with stepwise regression based on ANN for estimating S-system models of biological networks. They later (Voit E.O. 2004a) used the method of differentials substitution with slope estimates and system decoupling method of identification. With universal function computed by ANN in sequential and parallel order the inverse problems could be further simplified (BSTLab 2007)

Ebenbauer (Ebenbauer 2007) presented a dynamical (Lax) system that computes eigenvalues and diagonalizes matrices in real spectrum. The stability of the derived Lax system was determined by checking all the computed eigenvalues of the derived matrices. Fajarewicz et al. (Krzysztof Fajarewicz & Swierniak 2007) addressed the problem of fitting ODE models of cell signalling pathways. The fitting procedure, which is based on the generalized backpropagation through time GBPTT - an extension of backpropagation through time (BPTT) known in neural network theory - involved measurements taken at discrete time moments with concentrations of protein, protein complexes, and messenger RNAs (mRNAs) as time variables. This GBPTT, similar to other approaches used in signal flow graphs theory and electrical circuits, is a structural formulation of adjoint sensitivity analysis suitable for solving hybrid problems and may be used in constructing an ODE model of a signalling pathway. Chou, Voit, et al. (Chou I-Chun 2007) proposed a novel three-way alternating regression (3-AR) method and optimised the technique to estimate parameters from data using S-distributions. They used both real and noisy (*in silico*) data acquired from both S-distributions and traditional statistical distributions. This technique performed reasonably well and failed to converge only in very few cases. Derek Ruths et al. (Derek Ruths 2008) introduced the tool PathwayOracle which was developed in python to investigate and understand the dynamics of a signalling network using a simple, easy-to-build, unparameterized model with method capable of predicting signalling responses to experimental stimuli and conditions.

Barabassi et al. (Albert R. 2001) hinted on the interplay between network topology and

resilience (robustness against failures and attacks). They investigated and reported that the topology and evolution of real networks may be governed by robust organising principles. Most large complex networks are scale-free networks i.e. their degree distribution follows a power-law distribution. Hence modelling cellular systems should emphasise on capturing the network dynamics first before extracting topological features based on the philosophy that if the processes that assemble cellular networks are well captured, their topological features may be easily inferred. Torres et al. (Torres N.V. 2003) identified systemic level understanding of complex systems as a key requirement. A combination of structural and dynamic analyses of signalling networks is gradually becoming common practice (Papin J.A. 2005). Structural analysis deals with the identification of key pathways that determine the behaviour of a system (Andrea Sackmann & Koch 2006), while dynamic analysis involves predicting the changes in the concentration of signalling proteins over time. Sackmann et al. (Andrea Sackmann & Koch 2006) described a systematic approach for modelling signal transduction pathways using Petri nets (bipartite directed multi-graphs with a theory that provides a variety of established analysis techniques) and logical structures translated from biological phenomena. In focusing on building a discrete model of a signal transduction network without *a priori* knowledge about the actual kinetic parameters they first performed qualitative analysis of the pathway using Petri net theory before progressing towards quantitative analysis. Kitayama et al. (Kitayama T. 2006) identified the need to avoid using iterative procedures or calculations in identifying large-scale systems. They proposed a simplified power-law approach to modelling large-scale metabolic pathways from the Jacobian representation and steady-state flux profiles of the system under a wide range of perturbations of metabolite concentrations. Bansal et al. (Bansal M. 2007) in analysing both *in silico* and expression profile data compared several reverse engineering methods (i.e co-expression network and clustering algorithms, bayesian networks, and some ODE) highlighting the importance of efficient reverse engineering strategies in modelling gene regulatory interactions and the need to carry out adequate assessment of inference methods. They also proposed appropriating reverse engineering algorithms to data subsets. Nemenman et al. (Nemenman I. 2007) reconstructed metabolic networks from high-throughput metabolite profiling data using ARACNE, an algorithm developed for reverse engineering of transcriptional regulatory networks. They used synthetic data of model of red blood cell metabolism for the evaluation of the performance of their reverse engineering method. The performance

of the method on metabolic data was comparable to that on gene expression data. They also highlighted the usefulness of testing potential inference methods with benchmark data simulated from known networks. Goel, Chou and Voit (Goel G. 2008) developed and presented the dynamic flux estimation (DFE) as a methodological framework for estimating parameters of dynamic models of metabolic networks from time series data. This method integrates two distinct phases: a model-free and assumption-free estimation; and a model-based estimation. In the model-free phase model inconsistency between the data and alleged topology are addressed, while in the model-based phase the primary focus is on detecting and correcting ill-formulated mathematical functions. This approach is a significant improvement on all previously developed inference methods because it facilitates data consistency between model and experimental data. Vilela, Vinga, et al. (Vilela M. 2009) later presented the DFE framework with an improved and more robust strategy for constructing metabolic models from time-series data. Voit, Goel, Chou and Fonseca (Voit E.O. 2009) combined process-based and data-driven strategies in estimating metabolic pathways from different data sources complementary to DFE using the glycolytic pathway in *Lactococcus lactis* as example. Gennemark and Wedelin (Gennemark P. 2009) set some benchmarks to enable proficiency evaluation and comparisons between various ODE-based system identification methods using both simulated and real time series data of 40 test cases. Their ultimate goal was to solve all system identification challenges with their algorithm within hours and without the need for high computing power. The following fundamental issues are raised: unambiguously specification of reproducible ODE identification problems as mathematical optimisation problems and finding reasonably simple standard ways of representing a wide range of identification problems. The test challenges range from problems based on chemical rate equations to challenges frequently used in the development of optimisation algorithms.

In identifying some of the challenges associated with modelling approaches over the last decade, we learn that traditional modelling approaches may help explain some of the functions of the key components involved in biological systems. However, it may be necessary to avoid making any *a priori* assumptions about any underlying mechanisms involved within the systems, especially in *in vivo* systems (Voit 2000) where fast data-driven network reconstruction algorithms may be required. Therefore it is important to always keep in mind that data-driven modelling of individual parts of the overall systems should be studied in isolation but rather geared towards complementing integrated system modelling to support

system level understanding. To this aim we may ask what type of modelling strategy best answers the research questions to be addressed.

In dynamic modelling using time series data, we have identified a new approach which might prove highly effective and can help build computational tools to support integrated system modelling. Some of the requirements for our approach have been adopted from a set of important needs identified in carrying out the above-mentioned review of system identification methods used in deterministic modelling, including those issues of concerns highlighted by Voit (Voit 2000). These requirements may be categorised into the following specifications: instant capture (extraction) of network structure and system dynamics information from time series data; total avoidance of *a priori* assumptions about underlying processes during model construction; ensuring data consistency between the constructed model and experimental data inputs; optimisation of parameter estimation methods to eliminate redundant parameters in models; the need to develop or adopt non iterative system identification methods; the need to develop a generic method that can support automated ODE-based or BST-based reverse engineering geared towards system level understanding and multimodel and multi-scale integration; the need to (re)formulate a theoretical framework for general inverse problem solution; unsupervised model construction; mathematically convenient model structure; efficient error detection strategy for correcting large-scale models; and support for (inter-model) recasting techniques.

Our proposed reverse engineering framework is developed specifically to address all the above-mentioned needs targeting optimal utilisation of limited (extremely small e.g. 3 time points) time series data. We assume that nearly all existing inference algorithm require large quantities of time series data to work and none often guarantees data consistency and accurate reproducibility of experimental time series data under data limitation. Voit et al. (Voit E.O. 2004b) cautioned on common mistakes and potential hindrances to effective identification of biological system identification using *in vivo* time series data. These include algorithmic difficulties of nonlinear regression analysis, validity and consequences of incorrect *a priori* assumptions made in model design.

Table 3.1: Summary of ODE-based deterministic modelling methods

#	Method	Model/System	Reference	Date	Summary	Other comment
1		power-law	Sorribas and Cascante et al.	1994	systems identification	
2	power-law	S-system	Irvine	1988	systems modelling	
3	GA	power-law ODE	Savageau et al.	1996	power-law model refinement strategy	
4	Least-squares	S-system	Tominaga & Okamoto	1998	time consuming process	
		power-law	Hernandez-Bermejo et al.	1999/2000	power-law model derived by least-squares minimisation criterion	
5	GA	S-system	Akutsu et al.	2000	poorer performance;	
6	LP-based	S-system	Akutsu et al.	2000	more data better performance	
7	numerical method	S-system	Alves and Savageau	2000	mathematically controlled strategy	
8		BST	Voit and Radivoyevitch	2000	BST framework	
9	multilinear regression	power-law	Diaz-Sierra and Fairen	2001	nonlinear optimisation algorithm	
10	Mechanics	Complex systems	Barabassi et al.	2001	understanding complex networks	
11	ANN		Almeida.	2002	biological sequence analysis	
12	coexpression net. & clustering algo., bayesian net., ODE	<i>insilico</i> data	Bansal et al.	2002	inference task is very difficult	
13		BST	Voit	2002	BST review	
14	GA	S-system	Kikuchi et al.	2003	modelling of genetic networks	
15	evolution strategies	ODEs	Moles et al.	2003	modelling of biochemical pathways	
16	GA	S-system	Kikuchi et al.	2003	metabolic pathways modelling	
17		ODE	Torres et al.	2003	optimization of biochemical systems	
18	Neural network and data smoothing	S-system/ Stiff ODE	Almeida and Voit	2003	Modelling of biological models	
19	GA	2nd-order DE	Nyarko and Scitovski	2004	parameter identification problems	
20	optimised ES/MA	S-system	Spieth et al.	2004	modelling gene regulatory networks	
21	differential substitution, decoupling, and slope estimates	S-system	Almeida and Voit	2004	BST review	
22	multivariate linear regression	ODE	Veflingstad and Voit	2004	priming inference algorithm	
23		BST	Voit et al.	2004	system identification challenges	
24	Coevolutionary algorithm	S-system	Kimura et al.	2005	analysis of actual DNA microarray data	
25	Evolutionary method	S-system and DE	Noman and Iba	2005	inference of gene regulatory networks	
26	S-approximation	S-system	Tournier	2005	optimization of biochemical systems	
27		GMA	Polisetty, Voit, et al.	2005	metabolic system identification	
28		S-system	Voit	2005	smooth bistable system	
29	simulated annealing	S-system	Gonzalez et al.	2006	modelling cell signalling pathways	
30	info. criteria based fitness	S-system	Noman and Iba	2006	identification of network topology	
31	Petri nets design	Petri nets	Sackmann et al.	2006	modelling signal transduction pathways	
32	evolutionary algorithm	S-system ODE	Searson et al.	2006	S-systems with evolutionary algorithms	
33	linearisation	ODE/S-system	Kitayama et al.	2006	modelling large-scale metabolic pathways	
34	interval analysis	S-system	Tucker and Moulton	2006	modelling of a metabolic network	
35	Alternating regression	S-system	Chou, Voit, et al.	2006	parameter estimation method	
36	Eigenvalues	dynamical (Lax) system	Ebenbauer	2007	lax system stability	
37	GBPTT/BPTT	ODE	Fujarewicz et al.	2007	modelling cell signalling pathways	
38	ARACNE algorithm		Nemenman et al.	2007	modelling red blood cell metabolism	
39	improved fitness function		Noman and Iba	2007	accurate parameter estimation	
40		mass action	Srividhya et al.	2007	reconstructing pathways from time course data	
41	3-AR	S-system	Chou, Voit, et al.	2007	para. estimation using S-distributions	
42	PathwayOracle tool	unparameterised model	Derek Ruths et al.	2008	structural and dynamic analyses	
43	eigenvector optimisation	Decoupled S-system	Vilela, Chou, Vinga, Vasconcelos, Voit and Almeida	2008	automated reverse engineering	
44	dynamic flux estimation (DFE)	S-system	Goel, Chou and Voit	2008	assumption-free estimation	
45	lin-log	power-law	Rosario and voit	2008	modelling time series data	
46	identification algorithm	40 ODE benchmark systems	Gennemark and Wedelin	2009	reverse engineering benchmarks	
47	DFE	S-system	Vilela, Vinga, Maia, Voit and Almeida	2009	improved assumption-free estimation	
48	constrained optimisation	GMA	Ko, Voit and Wang	2009	flux connectivity relationships	
49	DFE		Voit, Goel, Chou and Fonseca	2009	estimating metabolic pathways	
50		BST	Chou and Voit	2009	BST and estimation methods	
51		S-system	Voit	2009	identification methods	
52		BST	Voit	2009	pathway, design and operation	
53		BST	Voit	2009	challenges of modelling	
54	NIR algorithm		Gregoretti et al.	2010	network identification	
55	lin-log estimation	Lotka-Volterra	Voit and Chou	2010	pathway modelling	
56			Voit and Kemp	2010	systems biology research	
57	B-Spline		Wang, Glover and Qian	2010	study of inference algorithms	

List of some system identification and parameter estimation techniques for analysing real and artificial time series data of biological networks.

3.3 Understanding BST as a modelling approach

The BST (Savageau 1969) is a modelling framework that has emerged over the last 40 years as being useful to modelling of gene regulatory networks and metabolic pathways. BST, also referred to as canonical modelling (Voit 1991), is based on rigorous mathematical foundation and theorems. In contrast to other traditional model representation such as the Michaelis-Menten rate law (Briggs G.E. 1925), BST is preferable because of the overwhelming difficulty in determining the kinetic constants and coefficients of a Michaelis-Menten process-based model and the relatively straightforward computations of eigenvalues, sensitivities, gains, and other key characteristics of a BST model (Voit 2013). terms of capturing and approximating systems processes and gaining insight from the analyses of both the model itself and data. Though we accept that there cannot be exact mathematical description of processes in nature and all laws of nature can only be approximations, but still very useful approximation may be derived from using them (Voit 2000). One of the advantages of modelling based on BST is that the modeller is able to primarily focus on higher-level questions that a model is addressing without minding and specifying enormous and often problematic details involved at the atomic level. In this way, hindrances normally posed due to redundant kinetic parameters in other traditional models may be avoided and cumbersome use of rational functions in parameter estimation.

3.3.1 Analytical convenience of BST

BST provides a convenient structure for analytical purposes. As indicated by Voit (Voit 2000), this modelling formalism has the following features: minimal level of assumptions is required compared to other traditional modelling approaches; every parameter is unique, well-defined and has a clear meaning; the model structure is analytically convenient for carrying out comparative studies of alternative models in mathematically controlled experiments. In addition the power-law representation features are particularly suited to address our modelling research questions, i.e. constructing and integrating jacobian and power-law models as an integrated solution to reverse engineering problems; have strong support for the analysis of large-scale networks; have many system identification and parameter estimation methods developed to support its framework.

3.3.2 Half system ODE representation

The half system is a form of BST which provides a complete aggregation of system's processes to single net terms which serves as an approximation of the production (synthesis) and degradation (depletion) of the molecular constituents within the system. Depending on the objectives of the modelling task the Half system model may be set up in such a way that a powerful and convenient tool for mimicking system dynamics and predicting future outcomes may be developed. Adopting a half system model as a nonlinear model facilitates system identification and parameter estimation challenges. This task practically involves recasting such half system model to a log-linear form and practically applying appropriate estimation techniques to infer a solution to the system of log-linear differential equations from time series data in a data-consistent way. For this reason, the half-system can also be called a "*Lin-Log*" model. The half system representation of dynamical systems is of the form:

$$\dot{X}_i = \alpha_i \cdot \prod_{j=1}^{n+m} (X_j^{g_{ij}})$$

where $i = 1 \dots n$; X_i are the dependent variables; n is the number of dependent state variables; m is the number of independent state variables; g_{ij} are called kinetic orders and quantify the overall net effect of X_j on the production (or degradation) of X_i ; α_i are called rate constants - they quantify the overall net turnover rate of the synthesis and depletion of X_i . A consideration with the Half system representation might be to reduce the model structure and number of kinetic parameters by eliminating all independent variables. Careful elimination of all redundant parameters will produce a most reduced form of the Half-system

$$\dot{X}_i = \alpha_{i_{new}} \cdot \prod_{j=1}^n (X_j^{g_{ij}}(t))$$

This would be the result if we aggregated all constant product $\prod_{j=n+1}^{n+m} (X_j^{g_{ij}})$ into the rate constants α_i , forming a new $[\alpha_{i_{new}}] = \alpha_i \cdot \prod_{j=n+1}^{n+m} (X_j^{g_{ij}})$. If we remove all independent variables and repeat the procedure for each dependent variable, a complete set of dependent variables and ODEs in matrix form would be created as:

$$\begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \\ \vdots \\ \dot{X}_n \end{bmatrix} = \begin{bmatrix} \alpha_1 \cdot \prod_{j=1}^n X_j^{g_{1j}} \\ \alpha_2 \cdot \prod_{j=1}^n X_j^{g_{2j}} \\ \vdots \\ \alpha_n \cdot \prod_{j=1}^n X_j^{g_{nj}} \end{bmatrix}$$

According to Voit (Voit 2000), systems of these types of equations are mathematically interesting but often inconvenient for carrying out steady-state analysis of biochemical systems - each power law term can not equate to 0. It is important to understand why such mathematically interesting representations may be unsuited for the analysis of biological systems. With keen interest, we explored the potential power of Half-system in combination with the convenient features of S-system for the analysis of biological data. We have discovered a simple mechanism by which Half-systems may be recast to a convenient structure such as the S-system power-law alternative. This simple strategy will be explained in the appendix. The important thing to note here is that though a Half-system may be generally inconvenient for analysing biological systems, with some minor adjustments they can be transformed into an effective modelling strategy for modelling biological data.

3.3.3 S-system ODE representation

The BST formulation

$$\dot{X}_i = V_i^+ - V_i^-$$

where the production rate function $V_i^+ = \alpha_i \cdot \prod_{j=1}^{n+m} (X_j^{g_{ij}})$ and degradation rate function $V_i^- = \beta_i \cdot \prod_{j=1}^{n+m} (X_j^{h_{ij}})$ is a product of power-law functions of all independent and dependent variables; multiplied by a rate constant that determines the speed of the process (Voit 2000). This type of representation that uses exactly two separate aggregate terms to represent the production and degradation rates of a variable is called S-system, where the S refers to synergism and saturation of the investigated system (Voit 2000). In BST the S-system representation of dynamical systems is of the form:

$$\dot{X}_i = \alpha_i \cdot \prod_{j=1}^{n+m} (X_j^{g_{ij}}) - \beta_i \cdot \prod_{j=1}^{n+m} (X_j^{h_{ij}})$$

where $i = 1 \dots n$; n is the number of dependent state variables; m is the number of independent state variables; g_{ij} and h_{ij} are called the kinetic orders that quantify the net effect of X_j on the production and degradation of X_i ; α_i and β_i are called rate constants. These two factors determine the net turnover rates of the synthesis and degradation of X_i . In reduced form (aggregating all independent variables into the new $\alpha_{i_{new}}$ and $\beta_{i_{new}}$ factors), the most reduced general representation

$$\dot{X}_i = \alpha_{i_{new}} \cdot \prod_{j=1}^n (X_j^{g_{ij}}) - \beta_{i_{new}} \cdot \prod_{j=1}^n (X_j^{h_{ij}})$$

may be used, i.e.

$$\begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \\ \vdots \\ \dot{X}_n \end{bmatrix} = \begin{bmatrix} \alpha_1 \cdot \prod_{j=1}^n X_j^{g_{1j}} - \beta_1 \cdot \prod_{j=1}^n X_j^{h_{1j}} \\ \alpha_2 \cdot \prod_{j=1}^n X_j^{g_{2j}} - \beta_2 \cdot \prod_{j=1}^n X_j^{h_{2j}} \\ \vdots \\ \alpha_n \cdot \prod_{j=1}^n X_j^{g_{nj}} - \beta_n \cdot \prod_{j=1}^n X_j^{h_{nj}} \end{bmatrix}$$

Note that constant factors α_i and β_i in the production and degradation terms for X_i respectively, may be ≥ 0 , but must not be < 0 . Since the influence of the variable X_j on the production and degradation of the variable X_i is determined by the negative, zero, or positive value of the kinetic orders g_{ij} and h_{ij} , respectively, therefore activating or inhibitory influences may be specified by assigning either a positive (activating) value or negative (inhibiting) value. A value of $g_{ij} = 0$ (or $h_{ij} = 0$) turns the term (multiplier) $X_j^{g_{ij}}$ (or $X_j^{h_{ij}}$) of V_i^+ (or V_i^-) to 1, i.e. X_j exerts no direct influence on the production (or degradation) of X_i .

3.3.4 The Generalised Mass Action (GMA) System

Another valid power-function representation in BST is the generalised mass action (GMA) description. This representation is also referred to as multinomial system ² (Voit 2000). The GMA representation is of the form

$$\dot{X}_i = V_{i1}^{\pm} + V_{i2}^{\pm} + V_{i3}^{\pm} + \dots + V_{in}^{\pm}$$

²Peschel and Mende, 1986

where each V_{i1}^{\pm} has the same meaning as either the production rate term $V_i^+ = \alpha_i \cdot \prod_{j=1}^{n+m} (X_j^{g_{ij}})$ or degradation rate term $V_i^- = \beta_i \cdot \prod_{j=1}^{n+m} (X_j^{h_{ij}})$. In the case of a GMA system, each equation may contain one, two, or more than two terms (Voit 2000).

3.4 Modelling based on BST

In every data-driven modelling challenge there are some key questions and analyses to be considered. Some of the key questions we have identified are discussed in this section. It is important to consider some key questions related to data: if the data is of time series type, one should ask should the model support backcasting or forecasting? We recommend that a good model should be able to facilitate both forecasting and backcasting. Backcasting deals with working backwards from a well defined time point to identify the transitions that may connect the previous states to that point in time. Backcasting may also be used as a technique for populating more data to a data set where necessary. However, it is essential that a verification process be developed and followed to determine how well the backcasting method works. One natural way of testing backcasting performance is by backcasting within the given set of all observed data and comparing the backcasting results with the real data that may have been recorded at those time points that lie within the backcasting time period.

It is also important that models be able to predict where the system might go from specific time points. The ability to do this depends on the predictive power and forecasting technique being used. After the predictive model has captured and formulated all key system functions, the resultant predictive model may be used to determine events whose outcomes are yet to happen or be observed. The goal of any forecasting technique is often to estimate the numeric value of a dependent variable of a model at some future time point.

Whether or not a model does support both backcasting or forecasting the modeller must develop and apply a system identification procedure to determine the appropriate model structure to make up the overall model architecture.

3.4.1 System identification and parameter estimation considerations

Often assumptions are made by the modeller in the specification of a network model. However, in dynamic modelling it may be required that a system identification strategy be developed and installed to prime model structure searching before parameter estimation. This priming of model structure can drastically improve parameter search space performance, or hinder it. System identification is one of the most important challenges that the modeller must tackle in dynamic modelling. By system identification, we mean both the method to build a mathematical model of dynamical systems from measured data and the architectural outcome (shape or structure) of the model itself. Effective methods that can optimise the desired outcome should be sought by the modeller. Such methods are very useful for reducing the parameter estimation burden that all network modelling processes must overcome.

On the other hand, parameter estimation of all model parameters is performed only after the complete model structure has been determined, calculated or assumed. The main objective of the estimation task is to find a set of parameter values which best describes the model in a way such that the model is capable of simulating the original data. Since system identification and parameter estimation problems are related, the modelling tasks may be classified into one major challenge. The outcome of system identification and parameter estimation often depend on both the method that is being used and sometimes the ingenuity of the modeller. Therefore, we think it is necessary to develop a modelling framework to support system identification and parameter estimation which will be based on data and such methods should be standardised (i.e. mathematically formalised on a theoretical notion that supports the capability to simulate or reproduce an original data input) where possible.

Handling changed data or dynamic data input

Another consideration is to determine how and when the constructed model may adapt dynamically to dynamic data (Voit 2004, Veflingstad S.R. 2004). Changed input may be intrinsically triggered or caused by external factors and therefore may result in perturbed outcomes that is significantly different from previous results. A most challenging changed input scenario that might have to be tackled is that of a complete change in data sets, e.g.

user intervening into new input specification because of new data that may be generated from new experiments. In such cases, if the system identification and parameter estimation procedure is not robust enough (i.e. unstable) in handling highly dynamic data a complete system identification and parameter estimation process might have to be repeated each time. Such repetitions may be time-consuming and expensive if the parameter estimation approach used is slow. Complete automation in system identification and parameter estimation process is required to adequately facilitate computational modelling and handle changed data input. In addition optimisation of the network inference strategy that is being used can enhance system performance. So it is important to think in terms of whether the appropriate model is being used, ensuring that the model being used is robust and can handle small-scale and large-scale data, and making sure that the constructed model is both optimised (data-consistent), i.e. is able to simulate (reproduce) by itself the original data that was supplied as input.

Stability analysis consideration

Eigenvalue analysis (Voit 2004, Veflingstad S.R. 2004) may be employed to systems of differential equations to determine whether or not a system would reach a steady state. If the calculated eigenvalue of one of the dependent variables of the model is positive, we may regard such system to be unstable. So it is important to think of how best to perform stability analysis. It is worth mentioning that without successful parameter estimation there cannot be meaningful stability analysis - accurate estimation of all parameters is essential. Effective modelling approaches are considered to be those that incorporate efficient parameter estimation with effective stability analysis methods - one that is able to strike a balance between parameter estimation and equilibrium theories. This is a justifiable means of assessing how one constructed model might be better or worse than another candidate solution (constructed model of the same target system).

The complication of steady state computation based on Half-system

If the determination of the steady state condition of the system is one of the key objectives of the modelling tasks then an appropriate model formalism such as S-systems may far outperform most others, including Half-systems. A complication may arise from the steady state computation of Half-system based model. This is because every dependent variable

should have at least one negative term in the set of ODEs; otherwise, the system may not reach a steady state. In my opinion, this is the main reason why experts such as Voit have categorically stated that Half-systems are generally inconvenient for the analysis of biochemical systems (Voit 2000). Not only were they right in pointing out Half-systems deficiency; a Half-system model is not good at all for the computation of steady states. This is because a Half-system model only consists of a product of powers (exponents) of variables that are often positive in the beginning, and continue to remain positive; even when their kinetic orders (exponents) are negative. Having said that, there is a simple remedy to this crippling effect of not being able to compute steady state - that is changing the rate constants to rate functions.

Advantages of power-law models

To briefly recount the key advantages of using a power-law model over an alternative, one might easily be tempted to say simplified computations, simplified optimisation algorithms, existence of efficient inference methods, and direct interpretability of parameters (Voit 2004, Veflingstad S.R. 2004).

3.4.2 Major challenges of BST model and methods

Chou and Voit (Chou I-Chun 2009) categorised some of the system identification and parameter estimation challenges as either data-related or model-related; of mathematical structure, and related to optimisation and support algorithms. Other BST based inference method challenges may include integration of ODE, smoothing overly noisy data, estimation of slopes from data, complexity of the inference work, constraining parameter search space, data preprocessing, detection and correction of model redundancy, etc.

As pointed out by Voit (Voit 2009), the primary goals of most modelling tasks could include allowance to extrapolate new situations and support accurate prediction for the understanding of pathway, design and operation. In agreement with Voit's suggestion (Voit 2009), in tackling the challenges of modelling one must identify speed, reliability, robustness and convenience as essential factors in dynamic modelling. In order to adequately tackle this challenge we may adopt the BST modelling approach, including those simple methods of estimating parameter values in other canonical models, e.g. the Lotka-Volterra and

linear-logarithmic models Voit and Chou (Voit E.O. 2010).

Finally, the modelling results must be interpreted in the context of the biological pathways of interest and related to the data input.

Chapter 4

Novel reverse engineering and network inference methods

4.1 Method 1: core method - jacobian approach

Here, we propose and present a novel computational, robust method (solution) for constructing underlying network of interactions purely from time series data. The method described in this chapter is completely data-driven and does not require any *a priori* information to be predetermined to infer successfully.

The construction of an interaction network from time series data may be recast as seeking to identify a mathematical model that relates a given time point to its successor, consistent with every time step and for all measurables. This model may be expressed in the form of a n by m matrix, for n time points and m measurables. For a given time point t , the mathematical model must relate a measured value x_i at t , $x_i(t)$, to its value in the subsequent time point $x_i(t + 1)$. Moreover, this relation is required for all $x(1, \dots, m)$ and for all $t(0, \dots, (n - 1))$. This mathematical model, referred to here as a transformation matrix, must be calculated from the data, i.e. the jacobian model and partial derivatives must be inferred from data. This process may be described as an *inverse problem*, where we must identify the transformation matrix solution to the system describing the time series data.

Finding a time-invariant matrix to the time series inverse problem may require that the number of time points provided be at least equal to the number of measured variables

+ 1. Whenever the available data set is limited in size, the number of time points is less than or equal to the number of measured variables, solution identification becomes more challenging.

The method proposed here requires a minimum of 3 time points to infer a network model that fits. As we show later, as the number of time points increases the network reconstruction process becomes more accurate. However, even with (only) 3 time points, we show that the method produces a data-consistent model of a given data set. A model is data-consistent if it is capable of reproducing exactly the original time series data that was used to infer the model.

Any time series data set can be described by a system of ODEs, whose variables are the measurables in the physical system. For example, a gene regulatory network with multiple genes can be represented by a set of nonlinear ordinary differential equations (ODE) with the expression level of each gene represented as a variable (E 2004). Reverse engineering of such networks, through gene expression data analysis and reconstruction of gene regulatory networks, involves revealing the underlying network of gene-gene interactions from the measured dataset of gene expression. This process may involve a form of mapping the observed data to a constructed and representative network model inferred from data using reverse engineering techniques (Chen L. 2009). With respect to the identification of the transformation matrix, we show below how variables in the ODEs model map to matrix elements and our algorithm operates on this matrix to reverse engineer interaction network topology and edge weightings. Solving this inverse problem has implications for data modelling in systems biology generally and in particular in personalised medicine through for example protein interaction network modelling.

The algorithm we are about to present searches for a solution to the inverse problem, and the algorithm identifies the same solution every time for a given problem. However, if the size of the available dataset is small, i.e., total number of time points < number of measured variables, other potential solutions may still exist. Our approach is different from other forms of optimisation algorithms such as genetic algorithms that are based on finding the best set of parameters within a search domain, because starting points for the domain search process are decided prior to parameter estimation. Here, no single parameter is fixed prior to estimation and no start point comes into play; the solution must be derived based on the experimental data supplied. Though some parameters can be set to fixed values,

our parameter estimation technique is purely based on and driven by experimental data, without any need for fixed parameters.

We outline how dynamic systems may be described by systems of ordinary differential equations, and present a matrix-based approximation to the solution. We then describe our algorithm to solve the inverse problem through identification of the values of elements in this matrix. We demonstrate the capability of the algorithm to reverse engineer interaction networks with a worked example, provide summary statistics on algorithm performance and comment on the uniqueness of the solutions discovered.

4.2 Ordinary differential equation systems

It is common to use ODE formalisms to model and analyse data in complex dynamical systems (E 2004). Sometimes, simple linear ODE models are sufficient representations of the system in order to identify essential relationships among network components based on time series data analysis. Solutions to systems of first-order ODE may involve matrix factorisation to approximate a matrix exponential (Liao J.C. 2003). Specifically, the solution of the homogenous equation

$$\frac{dX}{dt} = J.X(t) \quad (4.1)$$

where X is a column vector representing the dependent components of the system, t is a time variable and J , the relative rate of change of $X(t)$, is a matrix often referred to as acobian, is given by

$$X(t) = e^{J.t} X_0 \quad (4.2)$$

where $X(0) = X_0$, the initial state of the system for a system of linear differential equations, involving a transforming matrix J and the solution $[e^{J*t} * X]$, a function of (J, X, t) , where X is a column vector representing the dependent components of the system, and t is a time variable that represents a regular time interval between any two successive states, if J (the transformation matrix inferred from time series data) is derived such that it remains unchanged throughout the system (from the initial condition to steady-state, and after), then the model $\dot{X} = J * X$ may be said to be time-invariant because J is not dependent on time. In other words, the partial derivatives (elements) of the jacobian matrix

(J) do not change and are not functions of time.

4.2.1 Problem statement

Let X_t and X_{t+1} be the state vectors known from given time series data, and assume that the time interval t_c is regular and of small magnitude. Then $\dot{X} \approx \frac{X_{t+1}-X_t}{t_c}$, and the inverse problem is mathematically equivalent to

$$\frac{X_{t+1} - X_t}{t_c} \approx \dot{X} = J.X \quad (4.3)$$

There may be more than one J matrix that satisfies the above equation, so the primary objective of the reverse engineering strategy is to find the best J that describes - with least error - the transformation from any state X_t to the next state X_{t+1} .

4.2.2 Relative rate of change

J, the relative rate of change in $X(t)$, may be described as

$$J = \frac{\dot{X}}{X} \quad (4.4)$$

that is, *J is the absolute rate of change \dot{X} in relation to the present state value X .* Since it is true that $\int \frac{\dot{X}}{X} dt = \int \frac{1}{X} dX = \ln X + c_1$, therefore one may describe J in terms of $\int J dt = J.t + c_2 = \ln X + c_1 \rightarrow J = \frac{d(J.t)}{dt} = \frac{d(\ln X)}{dt}$, which means that *J itself is that relative rate of change, and it is equivalent to the rate of change in the logarithm of $X(t)$* (Gilbert 1988).

The jacobian can, in a sense, be thought as a representation of the rate and extent of change over time for any component and its temporal influence on other components. The jacobian matrix may thus be regarded as a construct for describing system dynamics. The mathematical significance of determined eigenvalues, or characteristic values, of the jacobian can help inform understanding of the behaviour of systems near a stationary point. By observing the eigenvalues of the jacobian matrix in a given neighbourhood, an indication of the stability of the system can be obtained. If the eigenvalues all have a negative real

part, the system is said to be stable (Burke J.V. 2003). A positive real part in any of the eigenvalues means the system may be unstable at that point, exhibiting large transient peaks (Burke J.V. 2003). Regarding the determinant of the jacobian matrix, the absolute value of the determinant of the (jacobian) transformation matrix indicates the overall rate of change of all the measured variables (Axler 1995). A necessary and sufficient condition for the jacobian matrix invertibility is that the magnitude of its determinant > 0 (Axler 1995).

The algebraic representation of an ODE-based model solution for a time series problem is simple and straightforward.

$$\begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \\ \vdots \\ \dot{X}_n \end{bmatrix} = \begin{bmatrix} \frac{\partial X_1}{\partial X_1} X_1 + \frac{\partial X_1}{\partial X_2} X_2 \dots + \frac{\partial X_1}{\partial X_m} X_m \\ \frac{\partial X_2}{\partial X_1} X_1 + \frac{\partial X_2}{\partial X_2} X_2 \dots + \frac{\partial X_2}{\partial X_m} X_m \\ \vdots \\ \frac{\partial X_m}{\partial X_1} X_1 + \frac{\partial X_m}{\partial X_2} X_2 \dots + \frac{\partial X_m}{\partial X_m} X_m \end{bmatrix} \quad (4.5)$$

And this is related to the eigenvectors and eigenvalues representation as follows

$$\begin{aligned} X_1(t) &= e^{\lambda_1 t} \cdot [v_{1,1}] \cdot [p_1] + e^{\lambda_2 t} \cdot [v_{1,2}] \cdot [p_2] + \dots + e^{\lambda_n t} \cdot [v_{1,m}] \cdot [p_m] \\ X_2(t) &= e^{\lambda_1 t} \cdot [v_{2,1}] \cdot [p_1] + e^{\lambda_2 t} \cdot [v_{2,2}] \cdot [p_2] + \dots + e^{\lambda_n t} \cdot [v_{2,m}] \cdot [p_m] \\ &\dots \quad \dots \quad \dots \\ X_n(t) &= e^{\lambda_1 t} \cdot [v_{n,1}] \cdot [p_1] + e^{\lambda_2 t} \cdot [v_{n,2}] \cdot [p_2] + \dots + e^{\lambda_m t} \cdot [v_{n,m}] \cdot [p_m] \end{aligned}$$

where the parameters λ_i and v_i represent the eigenvalues and eigenvectors, respectively; and the initial condition, $[X_1(0), X_2(0), \dots, X_n(0)]$, is favourably chosen, i.e., decomposed and approximated, as the linear combination of the eigenvectors of the jacobian matrix using the parameter set $[p_1, p_2, \dots, p_m]$. Each measurable X_i (component within a network) is a function based on the initial condition. The initial condition (first measurement) or any state vectors at any timepoint may be rewritten (decomposed) in a form such that the parameter set $[p_1, p_2, \dots, p_m]$ becomes fixed (Gilbert 1988).

4.2.3 Complex nonlinearity in systems of nonlinear ODEs

Since most complex systems are nonlinear in nature, nonlinear models are required to describe them fully. Moreover, some systems may require that second-order ODEs be used to formulate a sufficient description of their dynamics due to the complex nature of the processes involved. However, in such cases, the inference of the partial derivatives becomes a much more difficult task. In particular, the appropriate model structure needs to be identified, or at least estimated with a reasonable degree of confidence, before model parameters are estimated.

Here, we suggest that first-order linear ODE solutions based solely on time series data, i.e., with no assumptions made about network structure, may be sufficient to derive valuable information about the most important links among variables if good search algorithms for network inference are employed.

In some cases modellers predetermine network structure prior to parameter estimation, e.g. (Tsai K.Y. 2005). By keeping model structure fixed in this way, the inference procedure is restricted and identification of other solutions that do not adhere to the fixed structure constraint is not achieved. Such techniques themselves are thus limited since they are only successful when the constraint imposed on the system is met. Of course, the rationale for fixing the network topology is that it eases identification of potential solutions for underdetermined systems (Tsai K.Y. 2005). As a result, many modellers tend to assign some fixed values to subsets of the kinetic parameters to ease the fitting process. A good algorithm is one which can allow kinetic parameters to be fixed and at the same time does not require *a priori* assumptions to be made to find an accurate solution.

We propose the following algorithm, effected via a combination of pre-conditioning, regression, analytical techniques, half and/or S-system approaches based on time series data. Our rationales for proposing linear ODE solutions to solving inverse problems in time series are as follows:

1. Second-order differential equations can be recast into first-order (ordinary) differential equations;
2. Any nonlinear ODE can be cast or approximated into a power-law form called (S-systems or half systems);
3. Nonlinear half systems are equivalent to systems of log-linear differential equations;

4. Log-linear differential equations may be used using the same techniques for solving systems of linear differential equations.

Therefore, the development of a robust method for solving systems of linear differential equations is relevant to non-linear problems. Multiple regression and reverse engineering techniques, including the logarithmic inverse function, are important steps we have considered. Here, however, we limit the focus of the paper to time series analysis under systems of linear ordinary differential equations.

4.2.4 Methodology

In a system of linear differential equations an inverse problem may be written as in (4.4):

$$\dot{X} = J * X$$

where \dot{X} and X are known vectors of same length, and J is the unknown matrix that must be identified. Note that there is difference between a general system of n linear differential equations with unknown (jacobian) elements and a general system of n linear equations with unknown vectors. The latter is much simpler to solve due to the reduction in the number of unknown parameters. However, the formulation of the inverse problem remains the same in structure.

A general system of m linear equations with m unknown parameters is of the form

$$b = A.x \tag{4.6}$$

with the following matrix equations:

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \dots + a_{1m}x_m \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2m}x_m \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mm}x_m \end{bmatrix} \tag{4.7}$$

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ & & \ddots & \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \quad (4.8)$$

where $b=[b_1, b_2, \dots, b_m]^T$ are the constant terms, $a_{11}, a_{12}, \dots, a_{mm}$ are the coefficients of the system, and $x = [x_1, x_2, \dots, x_m]^T$ are the unknown parameters; note $[v]^T$ denotes the transposed vector $[v]$. Finding a solution to the system above involves searching for values in the x vector where all the m equations are simultaneously satisfied and valid.

A general system of m linear differential equations with an unknown $m \times m$ jacobian matrix J is of the form $\dot{X} = J.X$ and has the following matrix equation:

$$\begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \\ \vdots \\ \dot{X}_m \end{bmatrix} = \begin{bmatrix} \frac{\partial X_1}{\partial X_1} X_1 + \frac{\partial X_1}{\partial X_2} X_2 + \dots + \frac{\partial X_1}{\partial X_m} X_m \\ \frac{\partial X_2}{\partial X_1} X_1 + \frac{\partial X_2}{\partial X_2} X_2 + \dots + \frac{\partial X_2}{\partial X_m} X_m \\ \vdots \\ \frac{\partial X_m}{\partial X_1} X_1 + \frac{\partial X_m}{\partial X_2} X_2 + \dots + \frac{\partial X_m}{\partial X_m} X_m \end{bmatrix} = \begin{bmatrix} \frac{\partial X_1}{\partial X_1} & \frac{\partial X_1}{\partial X_2} & \dots & \frac{\partial X_1}{\partial X_m} \\ \frac{\partial X_2}{\partial X_1} & \frac{\partial X_2}{\partial X_2} & \dots & \frac{\partial X_2}{\partial X_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial X_m}{\partial X_1} & \frac{\partial X_m}{\partial X_2} & \dots & \frac{\partial X_m}{\partial X_m} \end{bmatrix} \cdot \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{bmatrix} \quad (4.9)$$

where $\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{bmatrix} = X(t)$ is a known state vector recognised as the t^{th} vector of X (not

X at time t), \dot{X} is the derivative vector which may be calculated from two known state vectors (X_t and X_{t+1}) as $\dot{X} \approx \frac{X_{t+1}-X_t}{t_c}$ where t_c is the interval of separation, and J is the unknown $m \times m$ jacobian matrix. Therefore, at least two state vectors of the same length are required to define an inverse problem in systems of linear differential equations. The following multi-state representation defines an inverse problem involving $n+1$ different states:

$$\begin{bmatrix} \dot{X}_{10} \dot{X}_{11} \dots \dot{X}_{1n-1} \\ \dot{X}_{20} \dot{X}_{21} \dots \dot{X}_{2n-1} \\ \vdots \\ \dot{X}_{m0} \dot{X}_{m1} \dots \dot{X}_{mn-1} \end{bmatrix} = \begin{bmatrix} \frac{\partial X_1}{\partial X_1} & \frac{\partial X_1}{\partial X_2} & \dots & \frac{\partial X_1}{\partial X_m} \\ \frac{\partial X_2}{\partial X_1} & \frac{\partial X_2}{\partial X_2} & \dots & \frac{\partial X_2}{\partial X_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial X_m}{\partial X_1} & \frac{\partial X_m}{\partial X_2} & \dots & \frac{\partial X_m}{\partial X_m} \end{bmatrix} \cdot \begin{bmatrix} X_{10} X_{11} \dots X_{1n-1} \\ X_{20} X_{21} \dots X_{2n-1} \\ \vdots \\ X_{m0} X_{m1} \dots X_{mn-1} \end{bmatrix} \quad (4.10)$$

which is equivalent to

$$\begin{bmatrix} \frac{X_{11}-X_{10}}{t_c} & \frac{X_{12}-X_{11}}{t_c} & \dots & \frac{X_{1n}-X_{1n-1}}{t_c} \\ \frac{X_{21}-X_{20}}{t_c} & \frac{X_{22}-X_{21}}{t_c} & \dots & \frac{X_{2n}-X_{2n-1}}{t_c} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{X_{m1}-X_{m0}}{t_c} & \frac{X_{m2}-X_{m1}}{t_c} & \dots & \frac{X_{mn}-X_{mn-1}}{t_c} \end{bmatrix} = \begin{bmatrix} \frac{\partial X_1}{\partial X_1} & \frac{\partial X_1}{\partial X_2} & \dots & \frac{\partial X_1}{\partial X_m} \\ \frac{\partial X_2}{\partial X_1} & \frac{\partial X_2}{\partial X_2} & \dots & \frac{\partial X_2}{\partial X_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial X_m}{\partial X_1} & \frac{\partial X_m}{\partial X_2} & \dots & \frac{\partial X_m}{\partial X_m} \end{bmatrix} \cdot \begin{bmatrix} X_{10}X_{11} \dots X_{1n-1} \\ X_{20}X_{21} \dots X_{2n-1} \\ \vdots \\ X_{m0}X_{m1} \dots X_{mn-1} \end{bmatrix} \quad (4.11)$$

assuming the states values are captured at regular time interval t_c . The smaller the value of t_c the better the outcome of this derivative approximation; this is because the linear approximation, $e^{t_c} \approx 1 + t_c$, of $e^{t_c} = 1 + t_c + \frac{t_c^2}{2!} + \frac{t_c^3}{3!} + \dots$ improves as t_c gets smaller and closer to 0 (Gilbert 1988). The solution to this system of linear differential equations:

$$\frac{\begin{bmatrix} (X_{11} - X_{10}) (X_{12} - X_{11}) \dots (X_{1n} - X_{1n-1}) \\ (X_{21} - X_{20}) (X_{22} - X_{21}) \dots (X_{2n} - X_{2n-1}) \\ \vdots \\ (X_{m1} - X_{m0}) (X_{m2} - X_{m1}) \dots (X_{mn} - X_{mn-1}) \end{bmatrix}}{t_c} = \frac{\begin{bmatrix} \frac{\partial X_1}{\partial X_1} & \frac{\partial X_1}{\partial X_2} & \dots & \frac{\partial X_1}{\partial X_m} \\ \frac{\partial X_2}{\partial X_1} & \frac{\partial X_2}{\partial X_2} & \dots & \frac{\partial X_2}{\partial X_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial X_m}{\partial X_1} & \frac{\partial X_m}{\partial X_2} & \dots & \frac{\partial X_m}{\partial X_m} \end{bmatrix}}{1} \cdot \begin{bmatrix} X_{10}X_{11} \dots X_{1n-1} \\ X_{20}X_{21} \dots X_{2n-1} \\ \vdots \\ X_{m0}X_{m1} \dots X_{mn-1} \end{bmatrix} \quad (4.12)$$

$\rightarrow \frac{X_{mn+1}-X_{mn}}{t_c} = J * X_{mn}$ is thus

$$\begin{bmatrix} X_{11}X_{12} \dots X_{1n} \\ X_{21}X_{22} \dots X_{2n} \\ \vdots \\ X_{m1}X_{m2} \dots X_{mn} \end{bmatrix} = \left[\exp \left(\begin{bmatrix} \frac{\partial X_1}{\partial X_1} & \frac{\partial X_1}{\partial X_2} & \dots & \frac{\partial X_1}{\partial X_m} \\ \frac{\partial X_2}{\partial X_1} & \frac{\partial X_2}{\partial X_2} & \dots & \frac{\partial X_2}{\partial X_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial X_m}{\partial X_1} & \frac{\partial X_m}{\partial X_2} & \dots & \frac{\partial X_m}{\partial X_m} \end{bmatrix} \cdot t_c \right) \right] * \begin{bmatrix} X_{10}X_{11} \dots X_{1n-1} \\ X_{20}X_{21} \dots X_{2n-1} \\ \vdots \\ X_{m0}X_{m1} \dots X_{mn-1} \end{bmatrix} \quad (4.13)$$

Consequently solving an inverse problem in a system of linear differential equations is equivalent to identifying J , $\frac{\partial(X_1, \dots, X_m)}{\partial(x_1, \dots, x_m)}$ that fits the data best. This requires optimal estimation of the partial derivatives (elements) of J ; calculating all the entries (parameters) of the matrix of all first-order partial derivatives of vector-valued functions of variable i with respect to another variable j , where X_i or X_j represents the variable function of component i or j , respectively. Using simple vari-

ables, the solution may be rewritten as: $X_{(N+1)} = E.X_{(N)} = e^{J.t_c}.X_{(N)}$ which is on the one hand a representation of system of linear equations, $X_{(N+1)} = E.X_{(N)}$, and on the other hand a direct interpretation of a system of nonlinear equations, $X_{(N+1)} = e^{J.t_c}.X_{(N)}$, implying that $E = e^{J.t_c}$. Consequently E is equivalent to the matrix exponential (function) of the matrix product $J.t_c$. The time constant t_c is easily calculated as the difference in time between T_1 at any state in X_{N+1} and T_0 its previous state in X_N . E can easily be approximated by our new regression techniques (described below), and these are variant forms of regression for ODE systems. Often regression analysis is used to understand the relation between two or more interrelated variables. In systems of linear differential equations, regression analysis can be used to infer causal relationships or transformations between the model variables and states. So in principle, *the state transformation (or the transposive regression) matrix, E, is the matrix exponential of $(J.t_c)$.*

From these definitions, an appropriate method must be applied to find the inverse solution. One which is worthy of mention is the logarithm of E, which in this case is the actual result $(J.t_c)$ being derived from E such that its exponential equals E.

Approximating a value of E may comprise using a regression technique having the steps:

1. Acquire time series data with the number of time points ≥ 3 ;
2. Preprocess the measured state values of one or more components of the system using matrix transposition;
3. Undertake regression analysis of the resultant data using a Moore-Penrose pseudoinverse technique;
4. Postprocess the resultant data using matrix retransposition;
5. Calculate the logarithmic inverse of the retransposed result through application of eigenvectors and eigenvalues techniques;
6. Scale down the resultant data by factor (magnitude) of the time interval used.

The implementation of the steps 2-6 is shown in Sections 4.2.5 and 4.2.5 below. Section 4.2.12 considers the (artificial) simulation of time series data from a predeter-

mined network model using data discretisation technique (step 1). The simulated data will be used to test and assess the proposed inference (reverse engineering) algorithm.

4.2.5 Transposive and repressive regression methods

In solving a system of linear differential equations, we define the solution to an inverse problem as one conditioned on the property:

$$\exp^{(J \star t_c)} * X(t) = X(t_{+1})$$

or simply:

$$E * X(t) = X(t_{+1})$$

where $E \approx \exp^{(J \star t_c)}$, the time interval t_c is assumed to be regular; J is unknown at this point and must be identified, and $X(t)$ and $X(t_{+1})$ are known arrays of column vectors, each a representation of system states at two successive time points, termed *before* and *after*. Here we present for the first time two new algorithms to derive J , namely:

1. (T) Transposive Regression Algorithm
2. (R) Repressive Regression Algorithm

Derivation of the (T) *Transposive* Regression Algorithm

Steps

1. $E_1 * X(\text{before}) = X(\text{after})$
2. $X(\text{before})^T * E_1^T = X(\text{after})^T$
3. $X(\text{before}) * X(\text{before})^T * E_1^T = X(\text{before}) * X(\text{after})^T$
4. $E_1^T = [X(\text{before}) * X(\text{before})^T]^{-1} * X(\text{before}) * X(\text{after})^T$
5. $E_1 = ([X(\text{before}) * X(\text{before})^T]^{-1} * X(\text{before}) * X(\text{after})^T)^T$

$$6. E_1 = X_{(after)} * X_{(before)}^T * ([X_{(before)} * X_{(before)}^T]^{-1})^T$$

In steps 1-2 recasting the problem by matrix transposition is essential, because each state is represented by a column-vector either in $X_{(before)}$ or $X_{(after)}$, where $X_{(before)}$ is an array of states before the transformation $X_{(before)} = [X(0) X(1) \dots X(t-1)]$, and $X_{(after)}$ is an array of states after the transformation $X_{(after)} = [X(1) X(2) \dots X(t)]$. Steps 3-4 illustrate an application of the Moore-Penrose pseudoinverse, a widely known type of matrix pseudoinverse, independently introduced by Moore (E.H 1920), Bjerhammar (Arne 1951), and Penrose (Roger 1955). Finally, in steps 5-6, retranspositions put E_1 in proper order.

Derivation of the (R) *Repressive Regression* Algorithm

Steps

1. $E_2 * X_{(before)} = X_{(after)}$
2. $E_2 * X_{(before)} - X_{(before)} = X_{(after)} - X_{(before)}$
3. $X_{(before)}^T * (E_2 - I)^T = (X_{(after)} - X_{(before)})^T$
4. $X_{(before)} * X_{(before)}^T * (E_2 - I)^T = X_{(before)} * (X_{(after)} - X_{(before)})^T$
5. $(E_2 - I)^T = [X_{(before)} * X_{(before)}^T]^{-1} * X_{(before)} * (X_{(after)} - X_{(before)})^T$
6. $(E_2 - I) = ([X_{(before)} * X_{(before)}^T]^{-1} * X_{(before)} * (X_{(after)} - X_{(before)})^T)^T$
7. $(E_2 - I) = (X_{(after)} - X_{(before)}) * X_{(before)}^T * ([X_{(before)} * X_{(before)}^T]^{-1})^T$
8. $E_2 = (X_{(after)} - X_{(before)}) * X_{(before)}^T * ([X_{(before)} * X_{(before)}^T]^{-1})^T + I$

Here, in step 1 we first repress the equation by subtracting $X_{(before)}$ from both sides of the equation. In steps 2-3 we recast the problem by matrix transposition. In steps 4-5 the Moore-Penrose pseudoinverse is applied. And the re-transposition step is introduced in steps 6-7. The identity matrix (I) is added to both sides of the equation to derive E_2 on the left hand side in step 8.

4.2.6 The search for the jacobian matrix solution

It is not difficult to calculate the jacobian matrix once either E_1 or E_2 is found. There are two different methods to consider:

1. using eigenvalues and eigenvectors;
2. using a new approximation technique, presented here for the first time.

The difficulty in finding J is in calculating the principal matrix logarithm of E_1 or E_2 ; that is, the exact inverse of $\exp(J * t_c)$.

$$\exp(J * t_c) = E_1 \approx E_2$$

$$J = \frac{\log m(E_1)}{t_c} \approx \frac{\log m(E_2)}{t_c}$$

where $\log m(\dots)$ represents the matrix logarithm function.

4.2.7 Application of eigenvalues and eigenvectors

Assuming that E_1 or E_2 is diagonalisable, the following method may be used to obtain the jacobian matrix from E_1 or E_2 . First we seek to find the matrix v of eigenvectors of E_1 or E_2 as appropriate, referred to here as E_m for convenience (for either case). Each column of v is an eigenvector of E_m . We then find eigenvectors of E_m from v and E_m as $eig_m = v^{-1} * E_m * v$. Next we replace each diagonal element of eig_m by its natural logarithm and calculate the natural logarithm of E_m as $\log m(E_m) = v * \log m(eig_m) * v^{-1}$. Finally, J is then calculated from $\log m(E_m)$ as follows:

$$J = \frac{\log m(E_m)}{t_c} \approx \text{real}\left(\frac{(v * \log m(eig_m) * v^{-1})}{t_c}\right) \quad (4.14)$$

Note, only real values for parameters of J are considered.

4.2.8 A new method for calculating matrix logarithmic inverse

It is generally known that $\exp^{(J*\nabla p)}$ is approximately equal to $I+(J*\nabla p)$ on condition that ∇p is a number and small enough (Gilbert 1988). Here, we will introduce a scaling factor μ to t_c in order to approximate ∇p such that $\nabla p \approx t_c * \mu$. We may calculate the value of J to be:

$$I + J * (t_c * \mu) = [\exp^{J*t_c}]^\mu \quad (4.15)$$

$$J * (t_c * \mu) = [\exp^{J*t_c}]^\mu - I \quad (4.16)$$

$$J = \frac{([\exp^{J*t_c}]^\mu - I)}{(t_c * \mu)} \quad (4.17)$$

$$J = \frac{(E_m^\mu - I)}{(t_c * \mu)} \quad (4.18)$$

By substituting a small value for μ in the above equation, e.g. $10^{-4} \leq \mu \leq 10^{-11}$, we may approximate J. Note, the smaller the magnitude of this value the better the result of approximating J becomes. However, care should be taken not to allow the magnitude of this value to be smaller than this range in order to ensure that it is not approximated to zero internally. Note this new approximation technique is equivalent, in terms of the solution obtained, to (4.14).

4.2.9 Linking jacobian matrices and network models

The inference of the transformation matrix J and all its partial derivatives from experimental data sometimes requires multivariate regression to be performed on experimental data with the aim of minimising the residual error between the model and the data, particularly in systems of linear and nonlinear ordinary differential equations (ODE). For example, using the following system of linear ordinary differential equations:

$$\begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \\ \dot{X}_3 \\ \dot{X}_4 \\ \dot{X}_5 \\ \dot{X}_6 \end{bmatrix} = \begin{bmatrix} \frac{\partial X_1}{\partial X_1} & \frac{\partial X_1}{\partial X_2} & \frac{\partial X_1}{\partial X_3} & \frac{\partial X_1}{\partial X_4} & 0 & \frac{\partial X_1}{\partial X_6} \\ 0 & \frac{\partial X_2}{\partial X_2} & \frac{\partial X_2}{\partial X_3} & \frac{\partial X_2}{\partial X_4} & 0 & \frac{\partial X_2}{\partial X_6} \\ 0 & 0 & \frac{\partial X_3}{\partial X_3} & \frac{\partial X_3}{\partial X_4} & \frac{\partial X_3}{\partial X_5} & 0 \\ 0 & 0 & 0 & \frac{\partial X_4}{\partial X_4} & 0 & \frac{\partial X_4}{\partial X_6} \\ \frac{\partial X_5}{\partial X_1} & 0 & 0 & \frac{\partial X_5}{\partial X_4} & \frac{\partial X_5}{\partial X_5} & \frac{\partial X_5}{\partial X_6} \\ 0 & 0 & 0 & 0 & 0 & \frac{\partial X_6}{\partial X_6} \end{bmatrix} * \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{bmatrix}$$

one can interpret the jacobian matrix to mean a direct representation of the network interaction and systems dynamics as indicated in Figure 1.

4.2.10 Results

4.2.11 Method validation

To test our approach for reconstructing network models from time series data, we use artificial data generated from known network models. These data are generated by simulating time series data from those models, and reconstructing the original network from the time series data alone with no knowledge of the generative model. Importantly, the original network is not provided to the reconstruction method - only the time series data, and this provides a source of independent test data not used in method construction. We test our method on 100 randomly generated networks as explained in 4.2.16. We consider two data discretisation approaches to generating time series data, namely:

1. discretisation using simple continuous models;
2. discretisation using application of eigenvectors and eigenvalues.

The objectives of the reverse engineering method and assessment are:

1. to simulated through discretisation of a continuous model (multivariate time series) data with known network models (jacobian matrices);

2. to ensure that the simulated data is noise-free (noiseless);
3. to ensure that important features of the data such as correlation between the variables are preserved;
4. to test and evaluate the performance of our inference algorithms based on minimum number of states (\leq number of measured variables+1).

The test methods avoid independent simulation of time series data of any single variable; only multivariate discretisation is used. To demonstrate the performance of our inference algorithms we also avoid using the integral function during the discretisation process. First we describe the two (2) discretisation techniques than can be used used to generate an experimental time series data from a single jacobian network model. We then present an expanded example of how to reverse engineer the original jacobian matrix (network model) purely from the simulated data. Finally, we present summary statistics of performance of the method for reconstruction of a large number of networks generated at random.

1

4.2.12 Data discretisation using a simple continuous model

The solution to a system of ordinary differential equation $\dot{X}(t) = JX(t)$ is $X(t) = e^{Jt}X_0$ as noted previously, so in order to discretise at regular time step intervals t (in this example our time step is 0.25 seconds), we define our discretisation process at any state k as :

$$X(k) = e^{Jt} \cdot [e^{J(k-1)t} \cdot X(0)] = e^{Jt} \cdot X(k-1) \quad (4.19)$$

¹ These subsections 4.2.12 and 4.2.12 should be viewed as two alternative approaches to simulate an artificial experimental time series data for testing a reverse engineering method. In those sections we introduced and demonstrated how to create a time series data set called DS. Please note that DS is not a matrix but rather is an array of state vectors generated at regular time points a sample of time series data. So the reverse engineering of the time series DS is expected to produce the jacobian matrix J because the data DS was simulated from the network model defined by J . We used these sections to first demonstrate how to generate one of many quantities of test time series data sets that were used during the simulation experiments that led to the development of the fundamental algorithms introduced in the thesis as Core methods. Each generation (instance) only requires a different jacobian matrix to be formulated. A matrix construction algorithm is developed and presented in an appendix section. Hope this information helps!

or at state (k+1):

$$X(k+1) = e^{Jt} \cdot [e^{J(k)t} \cdot X(0)] = e^{Jt} \cdot X(k) \quad (4.20)$$

where $X(0), X(k-1), X(k), X(k+1)$ are the vector values at the time points 0, k-1, k, and k+1, respectively.

Define a network model that is to be used to simulate a test time series data as having the jacobian transformation matrix

$$J = \begin{bmatrix} -0.19242 & -0.17738 & -0.80447 & -1.148 & 0 & 0.10009 \\ 0 & -0.19605 & 0.69662 & 0.10487 & 0 & -0.54453 \\ 0 & 0 & 0.83509 & 0.72225 & -0.43897 & 0 \\ 0 & 0 & 0 & 2.5855 & 0 & -0.60033 \\ -1.4224 & 0 & 0 & -0.66689 & 0.84038 & 0.48997 \\ 0 & 0 & 0 & 0 & 0 & 0.73936 \end{bmatrix}$$

and the initial condition, $X(0)$, as:

$$X(0) = \begin{bmatrix} 1.7119 \\ -0.19412 \\ -2.1384 \\ -0.83959 \\ 1.3546 \\ -1.0722 \end{bmatrix}$$

Therefore, the state vector $X(1) = e^{J \cdot t} * X(0)$ is:

$$X(1) = \begin{bmatrix} 2.4189 \\ -0.48906 \\ -2.9903 \\ -1.3564 \\ 0.90592 \\ -1.2898 \end{bmatrix}$$

where t=0.25 seconds. Likewise, state vector $X(6) = e^{J \cdot 6 \cdot t} * X(0)$ is then calculated

and the result is:

$$X(6) = \begin{bmatrix} 19.4655 \\ -6.0087 \\ -16.122 \\ -24.7886 \\ -10.9378 \\ -3.2502 \end{bmatrix}$$

Note that $X(6)$ could have been calculated from X_1 as: $X(6) = e^{J \cdot 5 \cdot t} * X(1)$. If we defined a time series dataset $DS_1 = [X(1) \ X(2) \ X(3) \ X(4) \ X(5) \ X(6)]$, i.e., a time series data set for the next six states at regular interval of 0.25 seconds after the initial condition, then DS would be:

$$DS = \begin{bmatrix} 2.4189 & 3.4924 & 5.1484 & 7.7765 & 12.0971 & 19.4655 \\ -0.48906 & -0.92226 & -1.5502 & -2.4683 & -3.8477 & -6.0087 \\ -2.9903 & -4.1059 & -5.6107 & -7.7393 & -10.9459 & -16.122 \\ -1.3564 & -2.293 & -4.0204 & -7.245 & -13.3126 & -24.7886 \\ 0.90592 & 0.1048 & -1.2125 & -3.269 & -6.3694 & -10.9378 \\ -1.2898 & -1.5517 & -1.8667 & -2.2457 & -2.7017 & -3.2502 \end{bmatrix}$$

Assuming that the data set DS is a given experimental time series data that must be reverse engineered, is it possible to infer the original jacobian transformation matrix J purely from the given experimental time series data (DS)? Yes, it may be possible if we use the Transposive regression method (TRM) as demonstrated in subsection 4.2.14.

4.2.13 Data discretisation using eigenvectors and eigenvalues

This section should be viewed as an alternative approach to the method described in 4.2.12. Here we aim to use introduce another approach by which the time series data set DS generated above may be generated. This is to demonstrate we have a good understanding of how to generate time series data sets and appropriate a generated time series data to the network model that can simulate such data. It is widely known that the solution set of the system of ordinary differential equations $\dot{X}(t) = J \cdot X(t)$ can be represented by any combination of exponential functions of eigenvalues and

their eigenvectors in the form:

$$\begin{bmatrix} X_1(t) \\ X_2(t) \\ \vdots \\ X_m(t) \end{bmatrix} = \begin{bmatrix} e^{\lambda_1 t} \cdot [v_{11}] \cdot [p_1] + e^{\lambda_2 t} \cdot [v_{12}] \cdot [p_2] + \dots + e^{\lambda_n t} \cdot [v_{1m}] \cdot [p_m] \\ e^{\lambda_1 t} \cdot [v_{21}] \cdot [p_1] + e^{\lambda_2 t} \cdot [v_{22}] \cdot [p_2] + \dots + e^{\lambda_n t} \cdot [v_{2m}] \cdot [p_m] \\ \vdots \\ e^{\lambda_1 t} \cdot [v_{m1}] \cdot [p_1] + e^{\lambda_2 t} \cdot [v_{m2}] \cdot [p_2] + \dots + e^{\lambda_n t} \cdot [v_{mm}] \cdot [p_m] \end{bmatrix}$$

where the initial condition, X_0 , is represented as a linear combination of the eigenvectors of J (Gilbert 1988). Through further analysis, we introduce the matrix form of the initial condition as

$$\begin{bmatrix} X_1(0) \\ X_2(0) \\ \vdots \\ X_m(0) \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1m} \\ v_{21} & v_{22} & \dots & v_{2m} \\ \dots & \dots & \dots & \dots \\ v_{m1} & v_{m2} & \dots & v_{mm} \end{bmatrix} \cdot \begin{bmatrix} p_1 & \dots & \dots & \dots \\ \dots & p_2 & \dots & \dots \\ \dots & \dots & \vdots & \dots \\ \dots & \dots & \dots & p_m \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

so that we might present our general matrix form solution to be

$$\begin{bmatrix} X_1(t) \\ X_2(t) \\ \vdots \\ X_m(t) \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1m} \\ v_{21} & v_{22} & \dots & v_{2m} \\ \dots & \dots & \dots & \dots \\ v_{m1} & v_{m2} & \dots & v_{mm} \end{bmatrix} \cdot \begin{bmatrix} p_1 & \dots & \dots & \dots \\ \dots & p_2 & \dots & \dots \\ \dots & \dots & \vdots & \dots \\ \dots & \dots & \dots & p_m \end{bmatrix} \cdot \begin{bmatrix} e^{\lambda_1 \cdot t} \\ e^{\lambda_2 \cdot t} \\ \vdots \\ e^{\lambda_m \cdot t} \end{bmatrix}$$

where each column vector in

$$\begin{bmatrix} v_{11} & v_{12} & \dots & v_{1m} \\ v_{21} & v_{22} & \dots & v_{2m} \\ \dots & \dots & \dots & \dots \\ v_{m1} & v_{m2} & \dots & v_{mm} \end{bmatrix}$$

is an eigenvector of the jacobian matrix J and the parameter set $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ is a set of the eigenvalues of J. The parameter e is used to denote the exponential of 1. Note that the parameter set $\{p_1, p_2, \dots, p_m\}$ can easily be calculated by regression

analysis (Gilbert 1988). Using the example in the previous section where $J =$

$$\begin{bmatrix} -0.19242 & -0.17738 & -0.80447 & -1.148 & 0 & 0.10009 \\ 0 & -0.19605 & 0.69662 & 0.10487 & 0 & -0.54453 \\ 0 & 0 & 0.83509 & 0.72225 & -0.43897 & 0 \\ 0 & 0 & 0 & 2.5855 & 0 & -0.60033 \\ -1.4224 & 0 & 0 & -0.66689 & 0.84038 & 0.48997 \\ 0 & 0 & 0 & 0 & 0 & 0.73936 \end{bmatrix}$$

the eigenvalues and eigenvectors of J are calculated to be: $\text{real}(\text{eigVec}) =$

$$\begin{bmatrix} 0.085886 & 0.085886 & -0.30478 & -0.30478 & -0.44332 & 0.14915 \\ -0.054869 & -0.054869 & 0.6724 & 0.6724 & 0.11344 & -0.56798 \\ 0.1344 & 0.1344 & -0.15878 & -0.15878 & 0.32878 & -0.27401 \\ 0 & 0 & 0 & 0 & 0.82485 & 0.21687 \\ -0.74638 & -0.74638 & -0.39895 & -0.39895 & 0.04612 & 0.29707 \\ 0 & 0 & 0 & 0 & 0 & 0.66692 \end{bmatrix}$$

$\text{imag}(\text{eigVec}) =$

$$\begin{bmatrix} 0.32489 & -0.32489 & 0.34644 & -0.34644 & 0 & 0 \\ -0.25758 & 0.25758 & 0 & 0 & 0 & 0 \\ -0.49251 & 0.49251 & 0.11919 & -0.11919 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.3693 & -0.3693 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

eigVec=

$$\begin{bmatrix} 0.0859 + 0.325i & 0.0859 - 0.325i & -0.30478 + 0.346i & -0.30478 - 0.346i & -0.443 & 0.14915 \\ -0.0549 - 0.258i & -0.0549 + 0.258i & 0.6724 & 0.6724 & 0.113 & -0.56798 \\ 0.1344 - 0.491i & 0.1344 + 0.491i & -0.15878 + 0.119i & -0.15878 - 0.119i & 0.328 & -0.27401 \\ 0 & 0 & 0 & 0 & 0.824 & 0.21687 \\ -0.74638 & -0.74638 & -0.39895 + 0.36i & -0.39895 - 0.36i & 0.046 & 0.29707 \\ 0 & 0 & 0 & 0 & 0 & 0.66692 \end{bmatrix}$$

eigVal =

$$\begin{bmatrix} 1.004 + 0.6191i & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.004 - 0.6191i & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.36055 + 0.1235i & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.36055 - 0.1235i & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.5855 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.73936 \end{bmatrix}$$

Setting the initial condition to

$$X(0) = \begin{bmatrix} X_1(0) \\ X_2(0) \\ X_3(0) \\ X_4(0) \\ X_5(0) \\ X_6(0) \end{bmatrix} = \begin{bmatrix} 1.7119 \\ -0.19412 \\ -2.1384 \\ -0.83959 \\ 1.3546 \\ -1.0722 \end{bmatrix} = eigVec * P * \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

implies that $P =$

$$\begin{bmatrix} -0.9804 - 2.2804i & 0 & 0 & 0 & 0 & 0 \\ 0 & -0.9804 + 2.2804i & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.020446 - 0.5584i & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.020446 + 0.5584i & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.59519 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1.6076 \end{bmatrix}$$

with this estimated parameter set found through regression analysis. Therefore, we define our second discretisation process at any time point t as :

$$X(t) = \begin{bmatrix} X_1(t) \\ X_2(t) \\ X_3(t) \\ X_4(t) \\ X_5(t) \\ X_6(t) \end{bmatrix} = eigVec * P * \begin{bmatrix} e^{(eigVal_{11}*t)} \\ e^{(eigVal_{22}*t)} \\ e^{(eigVal_{33}*t)} \\ e^{(eigVal_{44}*t)} \\ e^{(eigVal_{55}*t)} \\ e^{(eigVal_{66}*t)} \end{bmatrix}$$

Therefore, the timepoint at time $t \rightarrow 0.25$ secs becomes:

$$X(1) = \begin{bmatrix} X_1(0.25) \\ X_2(0.25) \\ X_3(0.25) \\ X_4(0.25) \\ X_5(0.25) \\ X_6(0.25) \end{bmatrix} = eigVec * P * \begin{bmatrix} e^{(eigVal_{11}*0.25)} \\ e^{(eigVal_{22}*0.25)} \\ e^{(eigVal_{33}*0.25)} \\ e^{(eigVal_{44}*0.25)} \\ e^{(eigVal_{55}*0.25)} \\ e^{(eigVal_{66}*0.25)} \end{bmatrix} = \begin{bmatrix} 2.4189 - 0.0i \\ -0.48906 \\ -2.9903 + 0.0i \\ -1.3564 \\ 0.90592 + 0.0i \\ -1.2898 \end{bmatrix}$$

We define time series dataset $DS_2 = [X_1 \ X_2 \ X_3 \ X_4 \ X_5 \ X_6]$ $DS_2 =$

$$\begin{bmatrix} 2.4189 - 0.0i & 3.4924 - 0.0i & 5.1484 - 0.0i & 7.7765 - 0.0i & 12.0971 - 0.0i & 19.4655 - 0.0i \\ -0.48906 & -0.92226 + 0.0i & -1.5502 + 0.0i & -2.4683 + 0.0i & -3.8477 + 0.0i & -6.0087 + 0.0i \\ -2.9903 + 0.0i & -4.1059 + 0.0i & -5.6107 + 0.0i & -7.7393 + 0.0i & -10.9459 + 0.0i & -16.122 + 0.0i \\ -1.3564 & -2.293 & -4.0204 & -7.245 & -13.3126 & -24.7886 \\ 0.90592 + 0.0i & 0.1048 + 0.0i & -1.2125 + 0.0i & -3.269 + 0.0i & -6.3694 + 0.0i & -10.9378 + 0.0i \\ -1.2898 & -1.5517 & -1.8667 & -2.2457 & -2.7017 & -3.2502 \end{bmatrix}$$

Assuming that the data set DS_2 is a given experimental time series data that must be reverse engineered, is it possible to infer the original jacobian transformation matrix J purely from the given experimental time series data (DS_2)? Yes, it may be possible if we use the Transposive regression method (TRM) as demonstrated in subsection 4.2.14.

4.2.14 Results: application of reverse engineering methods

We now demonstrate how well our reverse engineering (inverse problem solution) algorithms work using the time series created in the last section where

$$\begin{bmatrix} time(secs.) \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{bmatrix} = \begin{bmatrix} 0 & 0.25 & 0.5 & 0.75 & 1 & 1.25 & 1.5 \\ 1.7119 & 2.4189 & 3.4924 & 5.1484 & 7.7765 & 12.0971 & 19.4655 \\ -0.1941 & -0.48906 & -0.92226 & -1.5502 & -2.4683 & -3.8477 & -6.0087 \\ -2.1384 & -2.9903 & -4.1059 & -5.6107 & -7.7393 & -10.9459 & -16.122 \\ -0.8396 & -1.3564 & -2.293 & -4.0204 & -7.245 & -13.3126 & -24.7886 \\ 1.3546 & 0.90592 & 0.1048 & -1.2125 & -3.269 & -6.3694 & -10.9378 \\ -1.0722 & -1.2898 & -1.5517 & -1.8667 & -2.2457 & -2.7017 & -3.2502 \end{bmatrix}$$

and we define

$$X_{(before)} =$$

$$\begin{bmatrix} 1.7119 & 2.4189 & 3.4924 & 5.1484 & 7.7765 & 12.0971 \\ -0.19412 & -0.48906 & -0.92226 & -1.5502 & -2.4683 & -3.8477 \\ -2.1384 & -2.9903 & -4.1059 & -5.6107 & -7.7393 & -10.9459 \\ -0.83959 & -1.3564 & -2.293 & -4.0204 & -7.245 & -13.3126 \\ 1.3546 & 0.90592 & 0.1048 & -1.2125 & -3.269 & -6.3694 \\ -1.0722 & -1.2898 & -1.5517 & -1.8667 & -2.2457 & -2.7017 \end{bmatrix}$$

$$X_{(after)} =$$

$$\begin{bmatrix} 2.4189 & 3.4924 & 5.1484 & 7.7765 & 12.0971 & 19.4655 \\ -0.48906 & -0.92226 & -1.5502 & -2.4683 & -3.8477 & -6.0087 \\ -2.9903 & -4.1059 & -5.6107 & -7.7393 & -10.9459 & -16.122 \\ -1.3564 & -2.293 & -4.0204 & -7.245 & -13.3126 & -24.7886 \\ 0.90592 & 0.1048 & -1.2125 & -3.269 & -6.3694 & -10.9378 \\ -1.2898 & -1.5517 & -1.8667 & -2.2457 & -2.7017 & -3.2502 \end{bmatrix}$$

and $t_c = 0.25$.

We provide a pseudo code for the reverse engineering steps as follows.

Table 4.1: Reverse engineering method: pseudo code

Step	Description
1	$E * X_{(before)} = X_{(after)}$
2	$X_{(before)}^T * E^T = X_{(after)}^T$
3	$X_{(before)} * X_{(before)}^T * E^T = X_{(before)} * X_{(after)}^T$
4	$E^T = [X_{(before)} * X_{(before)}^T]^{-1} * X_{(before)} * X_{(after)}^T$
5	$E = ([X_{(before)} * X_{(before)}^T]^{-1} * X_{(before)} * X_{(after)}^T)^T$
6	$E = X_{(after)} * X_{(before)}^T * ([X_{(before)} * X_{(before)}^T]^{-1})^T$
7	Select $\mu : 10^{-12} < \mu < 10^{-6}$
8	Since $E^\mu = \exp^{(J * t_c * \mu)} \approx I + (J * t_c * \mu)$
9	Therefore $J \approx \frac{(E^\mu - I)}{(t_c * \mu)}$

In steps 1-2 the reverse engineering problem is stated and then reformulated

through matrix transposition. Steps 3-4 illustrate an application of the Moore-Penrose pseudoinverse, a widely known type of matrix pseudoinverse. In steps 5-6, retranspositions put E in proper order. In steps 7 a scaling factor μ is introduced to the product $J * t_c$ to scale down the product $t_c * \mu$ to satisfy the condition specified in step 8. Finally, the unknown jacobian matrix may be ‘reverse engineered’ using the approximation method derived in step 9 as a result of the conditioning requirement satisfied in step 8.

Example 1a: Application of the (\ddot{T}) Transposive Regression Algorithm

First calculate E_1 from $E_1 = X_{(after)} * X_{(before)}^T * ([X_{(before)} * X_{(before)}^T]^{-1})^T$

→

$$E_1 = \begin{bmatrix} 0.95158 & -0.042221 & -0.22248 & -0.42058 & 0.012662 & 0.059982 \\ 0.0012321 & 0.95216 & 0.18909 & 0.057435 & -0.010828 & -0.15008 \\ 0.022125 & -0.00028992 & 1.2306 & 0.28972 & -0.13526 & -0.027802 \\ 0.0000076294 & -0.000045776 & 0.0000076294 & 1.9086 & 0.0000019073 & -0.22946 \\ -0.38654 & 0.0082092 & 0.041029 & -0.18741 & 1.2323 & 0.15822 \\ -0.0000019073 & 0.0000038147 & 0.0000019073 & 0 & 0.00000023842 & 1.203 \end{bmatrix}$$

Then use either of the matrix logarithm techniques, introduced in Section 3.4, to reverse engineer J

$$J = \begin{bmatrix} -0.19248 & -0.17737 & -0.80446 & -1.148 & 0.0000042473 & 0.1001 \\ 0.0000067921 & -0.19605 & 0.69661 & 0.10488 & -0.00000047751 & -0.54451 \\ 0.000033966 & 0.000102 & 0.83509 & 0.72223 & -0.43896 & 0.00010752 \\ 0.000022578 & -0.00013168 & 0.000031848 & 2.5855 & 0.0000061946 & -0.60039 \\ -1.4224 & 0.000018331 & 0.00000762 & -0.6669 & 0.84038 & 0.49001 \\ -0.0000069427 & 0.000014074 & 0.0000044721 & -0.0000014319 & 0.000001096 & 0.73935 \end{bmatrix}$$

Example 1b: Application of the (R) Repressive Regression Algorithm

E_2 is calculated from $E_2 = (X_{(after)} - X_{(before)}) * X_{(before)}^T * ([X_{(before)} * X_{(before)}^T]^{-1})^T + I$.

→ $E_2 =$

$$\begin{bmatrix} 0.9516 & -0.042213 & -0.22248 & -0.42057 & 0.012661 & 0.059998 \\ 0.0012293 & 0.95216 & 0.1891 & 0.057434 & -0.010828 & -0.15009 \\ 0.022114 & -0.00031662 & 1.2306 & 0.28973 & -0.13527 & -0.027828 \\ 0.0000038147 & 0.000015259 & 0.0000019073 & 1.9086 & -0.00000047684 & -0.22943 \\ -0.38653 & 0.0082054 & 0.041027 & -0.18741 & 1.2323 & 0.15821 \\ 0 & -0.00000095367 & 0.00000023842 & 0.00000011921 & 0.00000011921 & 1.203 \end{bmatrix}$$

Reverse engineering J from E_2 then produces: J =

$$\begin{bmatrix} -0.19241 & -0.17731 & -0.80448 & -1.1479 & -0.0000046917 & 0.10017 \\ -0.00000044025 & -0.19605 & 0.69663 & 0.10487 & 0.00000066244 & -0.54452 \\ -0.0000064513 & -0.000016899 & 0.83509 & 0.72225 & -0.43896 & 0.0000048789 \\ 0.000010899 & 0.000044338 & 0.0000026542 & 2.5855 & -0.00000098578 & -0.6003 \\ -1.4223 & 0.000031253 & -0.0000021115 & -0.66688 & 0.84037 & 0.49 \\ 0.00000007547 & -0.0000035556 & 0.0000010714 & 0.00000031275 & 0.00000043957 & 0.73936 \end{bmatrix}$$

4.2.15 Performance evaluation of algorithms

Network structure identification

The performance of the two algorithms introduced in this paper are evaluated in terms of their ability to identify original (unseen) network structures. Before parameter estimation, model structures should be determined. The approximated network structure is easily derivable from the jacobian matrix (derivable from the zero and non-zero entries in the matrix representation of the system of ODEs), regardless of the magnitude of the parameter entries. Therefore, we simplify a weighted jacobian matrix into its Boolean form, revealing the network structure (model architecture) in terms of presence and absence of links. These structures are in form of matrices

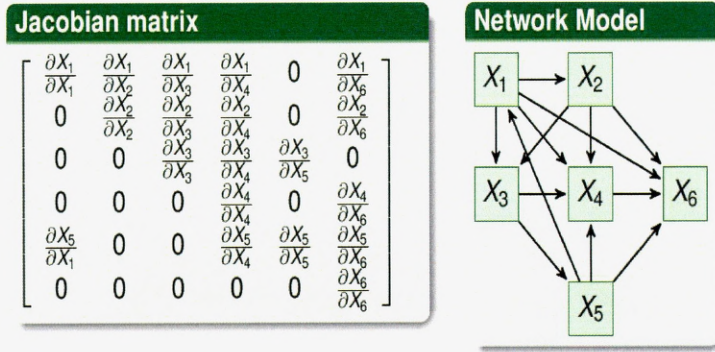


Figure 4-1: Relation between the derived boolean representation of the jacobian matrix and the corresponding network topology with inter-connected nodes.

containing only Boolean (0 or 1) entries: an entry of 1 depicts presence of an association (non-zero parameter in the jacobian matrix); otherwise 0 for no association. This initial approximation determines the network topology (see 4-1).

Network Connectivity Ratio

Given a (Boolean) matrix representation of the inferred jacobian matrix (network structure), we define network connectivity ratio as $\frac{NumOnes}{(NumOfOnes+NumOfZeroes)}$ where NumOfOnes and NumOfZeroes indicate the total number of 1s and 0s in the Boolean matrix respectively.

Metric criteria

Three key properties of network representation are used for performance evaluation:

- a) $\frac{Miss}{Total}$ ratio - a relative ratio of the number of missing correct links in the inferred matrix to the total number of links inferred;
- b) $\frac{Incorrect}{Total}$ relative ratio in terms of number of incorrect links in the inferred matrix to the total number of links inferred;
- and c) the norm score based on $\frac{Miss}{Total}$ and $\frac{Incorrect}{Total}$ ratios - an indication of the degree of closeness of the predicted network to the original network, measured as $\left(\left(\frac{Miss}{Total} \right)^2 + \left(\frac{Incorrect}{Total} \right)^2 \right)^{\frac{1}{2}}$.

Table 4.2: Summary statistics of algorithm performance

Test RunId.	Algorithm #	Size (No. of time points)	Network Connectivity (%)	$\frac{Miss}{Total}$ Ratio	$\frac{Incorrect}{Total}$ Ratio	Rank (Norm) Score
1	1	4	13	0.94167	0.14167	0.9575
	2	4	13	0.94167	0.14167	0.9575
2	1	5	16	0.63918	0.14145	0.65904
	2	5	16	0.63918	0.14145	0.65904
3	1	6	24	0.48424	0.33622	0.59018
	2	6	24	0.48424	0.33622	0.59018
4	1	7	23	0.34182	0.17576	0.38604
	2	7	23	0.34182	0.17576	0.38604
5	1	8	23	0.23455	0.097917	0.25519
	2	8	23	0.23455	0.097917	0.25519
6	1	9	23	0.14268	0.032794	0.14787
	2	9	23	0.14268	0.032794	0.14787
7	1	10	24	0.1063	0.043982	0.11522
	2	10	24	0.10797	0.03455	0.11377

Results of transposive regression (#1) and repressive regression (#2) algorithms for a range of numbers of time points for networks of 10 interacting components. 100 network instances each of seven (7) different datasets are simulated and tested using performance criteria established in 4.2.16.

4.2.16 Algorithm performance

We tested our algorithms on 700 simulated datasets with a range of numbers of time points (4-10) using the same data discretisation techniques specified in 4.2.12 and 4.2.13, all generated from networks with 25% connectivity ratio.

Well-formed jacobian matrices of artificial systems are required to generate data for performance evaluation of our method. A method for constructing nonsingular matrices was used to generate our artificial network data. Based on parameterisation of matrix implied determinants and minors, we were able to randomly generate a large set of new nonsingular jacobian matrices that were used to simulate test data by operating and manipulating products of factors of nonsingular matrices to guarantee that the jacobian matrices produced were not defective (see Appendix for nonsingular matrix construction). With hundreds of such matrices, we were able to generate a wide range of different artificial data to test our method. The number of non-zero elements in each of those matrices determined the network connectivity ratio for that system. Here, we fixed this to be 25% of the total parameters in a given matrix. Hence our matrices have 25% connectivity. Finally the results of the predicted net-

works were then compared to the corresponding information of the original networks recorded in the database.

Based on the performance evaluation criteria $\frac{miss}{total}$, $\frac{incorrect}{total}$ and the rank (norm) score, we analysed the results of the inferred network structure and measured deviation of network structure size in terms of network connectivity percentage. We assume that an inferred network structure has the potential to be a reasonable or good result if its connectivity ratio is between 20 and 30%. Table 1 shows that at least 60% of the number of dependent variables (here 6) is the minimum number of time points required to obtain a reasonable or good inference result. However, with our methods the reconstructed models are often data-consistent, that is, have the capacity to simulate or reproduce exactly the given dataset, irrespective of the number of time points (even when there are only a few time points). The evaluation criteria have been established to measure variation in performance depending on the number of time points. The results confirm that performance improves with an increase in the number of time points (a decrease in rank score indicates an increase in performance). The whole performance evaluation process is automated and does not need supervision or user intervention.

The network connectivity ratio indicates the estimated number of jacobian elements identified from (and used to explain) the available data set, which means that the richer the dataset, in terms of number of time points, the better the probability of identifying the correct links (or partial derivatives) that are suitable or valid, e.g. in Table 1, the first row indicated that on average at least 13 parameters (predicted from data to be non-zero) were ascertained to be valid parameters out of a total number of 25, whereas the last row showed that at least 24 parameters were identified as being valid. All (10) diagonal elements were included as valid entries by default.

Not every identified link is valid, although often the majority of them are. The network connectivity ratio reveals the complexity of the predicted networks, i.e. the total number of correct (and incorrect) links in the predicted network. The $\frac{Miss}{Total}$ value, as previously noted, reflects the number of correct links predicted.

As mentioned previously, the norm score based is based on the calculation $\left(\left(\frac{Miss}{Total} \right)^2 + \right.$

$\left(\frac{Incorrect}{Total}\right)^2\right)^{\frac{1}{2}}$. We assume that the lower the norm score (recorded in C.12) the better the performance of the inference method. A high valued norm score might still produce a data-consistent model, but the jacobian matrix of such system would be too sparse for the output structure to be a reasonable representation of the underlying data.

The result shows that performance, in terms of network structure identification, improves with an increase in the number of time points. The rank (norm) score (error ratio) value, a function of $\frac{Miss}{Total}$ and $\frac{Incorrect}{Total}$ ratios, confirms approximately $\frac{(1-0.115)}{1} * 100 = 88.5\%$ success rate in inference of network structure for datasets with size of 10 time points. It is remarkable that with such limited information on data and no information on topology, our inference methods are able to infer completely (100%) the actual network structure at times. Obviously, there is a critical point when the performance rank score is expected to be less than 0.5 failure threshold, i.e. when we might say that the combined effects of number of misses made or invalid predictions put together is lower than that of correct predictions as can be seen in the test runs labelled 4, 5, 6, and 7 with number of time points 7, 8, 9, and 10, respectively. The performances of the test runs labelled 1, 2, and 3 with number of time points 4, 5, 6, respectively, are below the success threshold due to insufficient data. As expected, under such extreme conditions of data inadequacy there would be far too small number of equations to be solved than number of parameter variables required to be estimated to identify all necessary and valid interactions. However, in many occasions our algorithms still, though constrained within the limits of those underdetermined test conditions, did not compromise in reducing $\frac{Incorrect}{Total}$ fraction to absolute minimum as possible. For instance, the results confirm that on average only about 14.2% of total predictions made is incorrect, even with limited number of 4 and 5 time points. The combination of pleasant outcomes such as those mentioned and the often guaranteed data-consistency in the data simulated from those inferred models, even under extreme lack of data, is one major strength common in both algorithms.

In summary, the results show an improvement in method performance with in-

creasing data size. Overall performance is close to optimum, i.e the original network is recovered with approximately 88.5% success rate on average, when the number of time points available is equal to number of measured variables even though this network is unseen to the algorithm and there are many possible data-consistent weighted networks. The main challenge is in keeping both the $\frac{Miss}{Total}$ and $\frac{Incorrect}{Total}$ ratios as low as possible. These two metrics may also be used in robustness and sensitivity analysis of any proposed method, keeping in mind that the primary objective of any proposed method is to minimise $\frac{Miss}{Total}$ and $\frac{Incorrect}{Total}$ values. Ultimately, the challenge is to preserve data-consistent model generation while at the same time maximising the likelihood of identifying the original set of links by inference.

4.2.17 Conclusion: jacobian method

Clearly, network structure identification and parameter estimation of dynamical systems are necessary steps in representing system dynamics in terms interaction networks. We demonstrate that algorithmic analysis of time series data may produce data-consistent models. On a promising note, the novel inference algorithms presented in this paper are identified, through simulation study and assessment, to develop such data-consistent models. As demonstrated in this thesis using a worked example, there is a strong theoretical basis for their use in time series data analysis. Moreover, their utility is demonstrated by their performance result under testing conditions using artificial data sets generated in silico.

We assessed the performance of our unsupervised inference algorithm using 700 hundred test networks, that is networks that were used to generate randomly valued, independent test data, through two different methods and similar in form (but not values) to the worked example, and importantly the underlying network structure was never provided to the algorithm in this validation. We demonstrated significant improvement in network reconstruction as more data became available, here increasing time series time points from 3 to 10, and showed very good performance as the data size tends to 10 time points. We recognise that 3 data points is a very small data set, but show that even with this limited time series data we are able to reconstruct

a data-consistent network. Our algorithms are aimed at simplifying and standardising the methods of finding unique solutions whenever they exist and using those standardised methods to adequately find other potential data-consistent solutions in non-unique scenarios. Of course, as the number of time points in the time series data supplied reduces, the number of possible networks able to explain the data increases. Note that this ability to work with limited data can be combined with the capacity for the approach to include a priori knowledge, and this knowledge may substantially reduce the solution space. Consequently the approach can blend available knowledge with knowledge gaps to produce data-consistent models of system dynamics.

Chapter 5

Dynamic modelling of DNA-damage response (DDR) pathways

An important challenge in cancer biology is the understanding of the ATM DNA damage response pathway, its regulation and dynamical behaviour in both normal and cancer cell lines. In this chapter we use a set of time series data from controlled experiments to investigate this pathway and seek to contribute to the interpretation of these experimental results using the inference methods described. The illustration presented here primarily is aimed at providing useful and practicable descriptions on how to predict disease nature and states using only time series data.

Experimentally we adopt a time dependent treatment strategy to investigate the signalling alterations and phospho induction of ATM and its substrates by both lower ($0.1\mu\text{M}$) and higher ($0.4\mu\text{M}$) concentration of Doxorubicin (Dox), a radiomimetic drug, with and without treatment of KU55933, an ATM kinase inhibitor. This drug intervention strategy which is based on varying levels of drug dosages is used to stratify experimental trials to provided rich quantitative data for information extraction purposes and knowledge discovery. The mathematical and computational modelling strategies assessed previously using artificially controlled *in-silico* experiments are now applied to real biological data. In this way, those same strategies that worked

during *in-silico* experimentation are now applied here in analysing time series data to extract new information about cell sensitisation and identify potential cancer drug therapies or inform the design of new therapeutic targets and alternatives in cancer treatments.

Assuming that quantitative data provide a representation of the unknown dynamics of signal activation of different proteins involved in DNA damage response pathway is available, it follows that such data may be analysed to generate new and useful information, e.g. hypotheses, about the possible or potential biological signalling, if extracted carefully. Furthermore, we demonstrate that by using the developed inference methods informative and data-consistent network models may be inferred and constructed from such experimental time series. The inference method applied is demonstrated to be useful in interpolating and extrapolating graphs of experimental data. In this case study, the dynamics of the predicted (interpolated and extrapolated) data are expected to be consistent with the original experimental data supplied.

5.1 Understanding the DNA-damage response pathway

How is ATM-mediated signalling different from ATR-dependent signalling? ATM is known to be activated in response to DNA damage and double-stranded DNA breaks. ATR on the other hand may initiate a signal cascade that results into cell cycle arrest. Chk1 and Chk2 are responsible for blocking entry into mitosis by (indirect) inhibition of Cdc25 activity. BRCA1, in addition to other control mechanisms, is responsible for rapid mobilisation of repair proteins to DNA damage. E2F is responsible for the switching on of important genes that encode proteins required for entry into the S-phase of the cell cycle. Histone H2AX is a repair component.

5.2 Aims and objectives

Here, using only experimental data of real biological experiments (subsection 5.2.2) obtained from the DNA-damage response (DDR) pathway we ask if it is possible to use this data to understand the roles of ATM and ATR and various mechanisms by which these pathways are being regulated and influencing the control of cellular responses to varying levels of DNA-damage in ATM-mediated and ATR-dependent signalling. In seeking answers to these questions the following objectives are set:

1. To study the ATM-mediated DNA-damage response pathway and determine its regulation in response to drug treatments with and without ATM inhibition. Since ATM is a central mediator of responses to DNA double-strand breaks (DSB) in cells, any insights into the mechanisms involved in ATM-related pathways or development of new sensitisers to therapeutic treatments may help improve existing benefits to cancer patients (M.B 2008).
2. To apply computational and mathematical modelling strategies to time series data of biologically controlled experiments and varying drug interventions to predict and explain cellular states and responses.
3. To perform *in silico* prediction of effects of ATM inhibition and activation kinetics in DNA damage response pathway
4. To interpret the results of (3) in order to understand the potential roles and mechanism of action of Doxorubicin (Dox) and its application in the treatment of cancer and inform the design of real biological experiments to understand the temporal effects and final outcomes of drug treatments based on varying levels of Dox input.

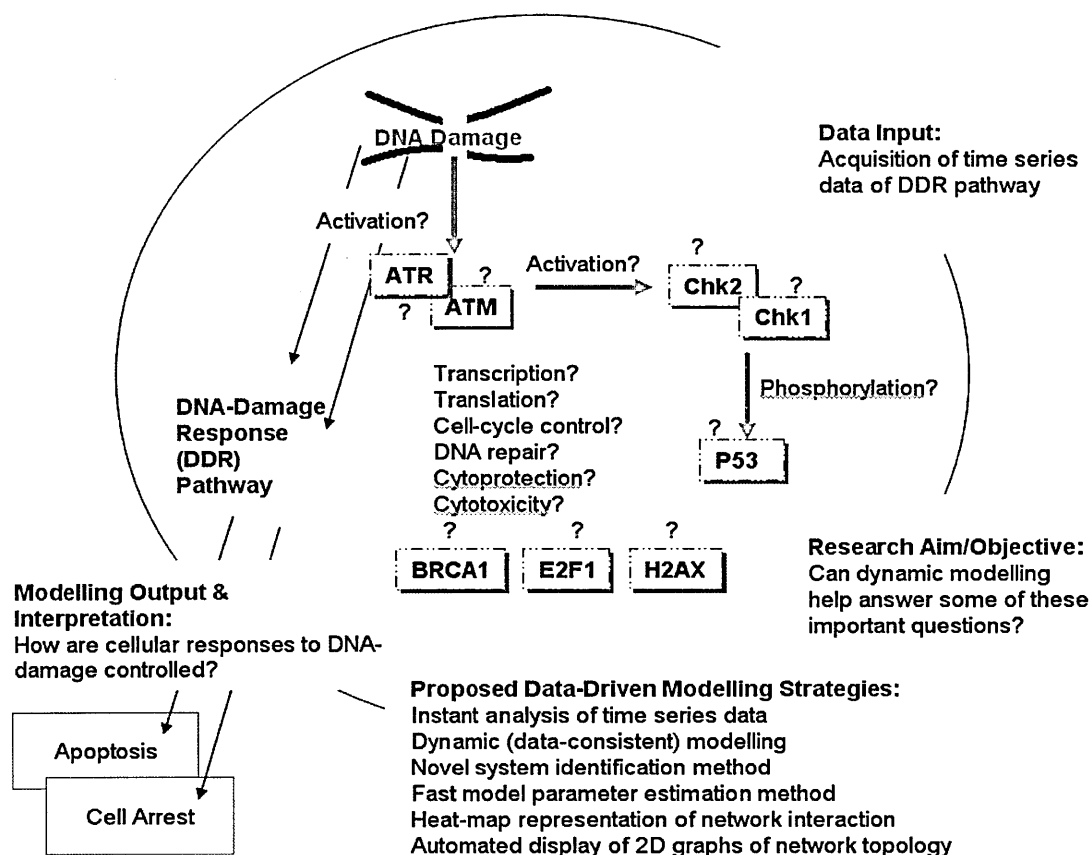


Figure 5-1: Understanding cellular responses to DNA-damage response pathways and ATM as a mediator of responses to DNA-damage.

5.2.1 DDR cancer biology

DNA damage may be caused by genotoxic agents, e.g. exogenous factors such as exposure to ionising radiation, UV light and some genotoxic chemicals. Certain internal factors such as cellular metabolism (instability) and replication errors may lead to abnormal consequences that may also result into DNA damage in cells (Date 2003). This damage in cells is often characterised by multiple signalling cascades (both known and unknown) that need to be understood.

It is generally known that X-ray-induced DNA damage may initiate a signalling pathway through the activation of the protein kinases ATM and ATR, which phosphorylate the kinases Chk1 and Chk2. The kinases Chk1 and Chk2 may also phosphorylate

other important proteins such as the regulatory protein p53. In other words, DNA damage may result into p53 activation. This decreased p53 degradation, which is a consequence of p53 phosphorylation, may result in an increase in p53 concentration (Alberts Bruce 2009). Since some cancers may be characterised by loss of p53 function and p53 plays an important role in cell-cycle arrest, it is important that we understand some of the key roles that ATM, ATR and p53 may play in ATM signalling and DNA damage response pathway. Also since Chk1 and Chk2 are particularly important in their roles in blocking cell cycle progression through inhibition of Cdc25 activity, understanding the various mechanisms involved before and after their phosphorylation by the ATM and ATR kinases associated with the site of DNA damage is also important.

We propose that by analysing and modelling quantitative time series data of pATM, pATR, p53, pChk1, pChk2, pBRCA1, pE2F1, pH2AX involved in ATM/ATR signalling pathway, new therapeutic targets may be identified for the development of effective cancer drugs and such analysis may be used to characterise the mechanism and dynamics of DNA damage response signalling.

The real biological experiments being conducted are designed to understand the temporal effects and final outcomes of drug treatments, and are based on varying levels of Doxorubicin (Dox) input. Dox is a DNA structural distortion inducer that damages DNA and can be used as a chemotherapeutic agent. In these experiments, the application of Dox is to be considered in two different dosages and in parallel over a fixed time period, under certain test conditions, i.e. in the presence or absence of KU55933 (KU), a widely known inhibitor of ATM.

As indicated earlier our objective is to seek to uncover and understand the underlying mechanisms by which DNA-damage response pathway may be impacted on, comparing both Dox and Dox+KU application methods. Here, these data are modelled to infer networks of interactions. These networks of interactions inferred from real experimental data may in some special cases require some additional biological experts' knowledge, particular in interpreting the modelling results to arrive at a concrete conclusion within a given biological context. However, it is essential that the

modelling process does not depend on additional input beyond those supplied time series data.

5.2.2 Biological experiments and method

The biological experiments were conducted by the biological domain expert and the measured time series data supplied to us in raw (tabular) form. An immortalised human keratinocyte cell line (HaCat) is used to generate quantitative time series data of DDR pathway dynamics following drug intervention for the development of an experimentally based deterministic model. Both a lower ($0.1\mu\text{M}$) and a higher ($0.1\mu\text{M}$) concentration of the widely used radiomimetic drug, Doxorubicin (Dox), was used for cell treatments with and without $10\mu\text{M}$ of ATM kinase inhibitor, KU55933 (Ku). The two concentrations of genotoxic agent were chosen in order to delineate the corresponding signalling dynamics at a lower and a higher degree of DNA damage and characterise the signalling alterations upon a repairable DNA damage and irreparable state (apoptosis). The time series experiments under different conditions (Table 1) were performed targeting all major proteins that respond to DNA damage (Table 2) and carried out their semi quantitative analysis. Relative quantifications of phospho induction of each protein were performed by high throughput HRP based ELISA to produce experimentally determined data consistent profile of the dynamic processes within biological system and parallel cytotoxicity analysis to provide kinetic parameters for a systems biology application.

The developed inference procedure requires that time series measurements be recorded at regular intervals. Either we ignore the 2-hr data or use an appropriate technique to first consider time intervals of 2 (rather than 4) hrs and then interpolate across the whole data set to determine data points every two hours or select the time series data values at regular time intervals of 4 hours. Therefore omitting the 2 hr data point is preferred over interpolating the data set mainly because we want to avoid any form of bias being introduced into the original data set. We note that the almost redundant 2-hr time point data is not removed from the original data set in order to use it for verification purposes after system identification procedure has been

Table 5.1: Treatment conditions for Neutral Red (NR) uptake based cell cytotoxicity assay.

Cell cytotoxicity assay (NR-uptake) conditions				
Time point	100nM Dox	100nM Dox +10 μ M KU	400nM Dox	400nM Dox +10 μ M KU
0	UT	UT	UT	UT
1	2 hr	2 hr	2 hr	2 hr
2	4 hr	4 hr	4 hr	4 hr
3	8 hr	8 hr	8 hr	8 hr
4	12 hr	12 hr	12 hr	12 hr
5	16 hr	16 hr	16 hr	16 hr
6	20 hr	20 hr	20 hr	20 hr

Table 5.2: DDR substrates analysed in the study

DDR Kinases	DDR substrates
pATM Serine 1981	pP53 S 15
pATR Serine 428	pBRCA1 S 1524
pChk2 Threonine 68	E2F1
pChk1 Serine 296	H2AX

executed.

5.3 Dynamic modelling and system identification methods

The biological experiments considered in this case study focus on the analysis of the DDR pathway data to understand the effects of DNA-damage and cellular responses to Dox input. Multiple key proteins of the DNA damage response pathway are therefore considered together with their activation levels. These activation levels are recorded by measuring fold phospho-induction of ATM, ATR, Chk1, Chk2, p53, E2F1, BRCA1, and H2AX producing experimental data of the DNA damage response pathway through cellular immunostaining with cytotoxicity assays.

We seek to deduce the effects of different drug treatments on cell death and cell survival by inferring underlying signalling networks from available quantitative data of

the DDR pathway. Our main objective is to determine how resulting network topologies inferred from data might help understand some of the key molecular mechanisms and activity impacted on by observing their perturbation differences in network architectures. Cell death responses to treatments with ATM inhibition are studied in parallel to identify new therapeutic targets. Modelling results should help gain new insights into how cellular responses are being regulated in response to ATM inhibition and drug input such as Dox targeting the DDR pathway.

5.3.1 DDR modelling challenge

Computational and mathematical modelling has many advantages. However, real practicable demonstration and (re)utilisation of the overall modelling process can be difficult to evidence. In our approach, we seek to find convincing evidence to demonstrate the importance of mathematical modelling and its application in *in-silico* analysis of biological data. To this aim we apply the technique introduced in Chapter 4 to model real time series data of ATM/ATR signalling pathway. An assessment of the entire modelling process will demonstrate whether the method is effective and capable of giving new insights into the underlying mechanisms that emerge as a result of drug intervention.

5.3.2 Computational modelling objectives

The following are points considered as part of the modelling objectives:

1. Heatmap images of experimental data sets should be instantly generated from the supplied biological data;
2. The data-consistent jacobian model constructed *must* be inferred purely from the given time series data;
3. The supplied time series data may require more data to be interpolated or extrapolated, e.g. more data may be required to generate a rich 2D plotted graph of system's states, therefore the inferential procedure *must* support backcasting and forecasting of unknown system states;

4. Extraction of either the overall or transient maps of network topology representative of the entire or some shorter period covered by the time series data.

5.3.3 An overview of the computational modelling approach

The models that will be used to describe the various systems will be purely data-driven. First, the given time series data will be analysed to produce a heatmap visual representation of all time series data input to the system. This is a particularly useful visual means of validating that the right set of data is being used. Such heatmaps may also serve other multi-purposes including effective aids for visualising both the data input and simulated data output as well as comparing and stratifying the magnitudes of all model parameters.

Figure 5-2 shows the time series data sets representing the effects of doxorubicin-induced mutation with or without ATM inhibition by KU. Time series measurements at lower ($0.1\mu\text{M}$) and higher ($0.4\mu\text{M}$) dose-intensities of both doxorubicin and doxorubicin+KU treatments are supplied. At $0.1\mu\text{M}$ Dox input a gradual and steady rise is observed in both pATM and pP_{53} up to 20 hours. However, a rise in pATM (up to 16 hours) and dramatic rise in pP_{53} levels is observed at $0.4\mu\text{M}$ Dox input. More inhibition of ATM by KU is observed at lower DNA-damage input. Higher activation of ATR in response to higher DNA-damage input is observed in the presence of KU.

The modelling framework in figure 5-3 illustrates the modelling approach adopted in this case study. This strategy shows how to analyse time series data of DNA damage response pathway to infer and construct a data-consistent predictive model of the DDR system. With such a constructed model new or missing time point data may be generated for forecasting or backcasting purposes. An unknown underlying network of interactions may be inferred from it without requiring *a priori* information about a specific part of the real network topology to be used. The constructed mathematical model could reveal new and diverse mechanisms about the underlying complex signalling networks depending on the treatment options and set of time points specified. Further studies may be required to confirm or determine the extent to which the modelling results or interpretation is conclusive.

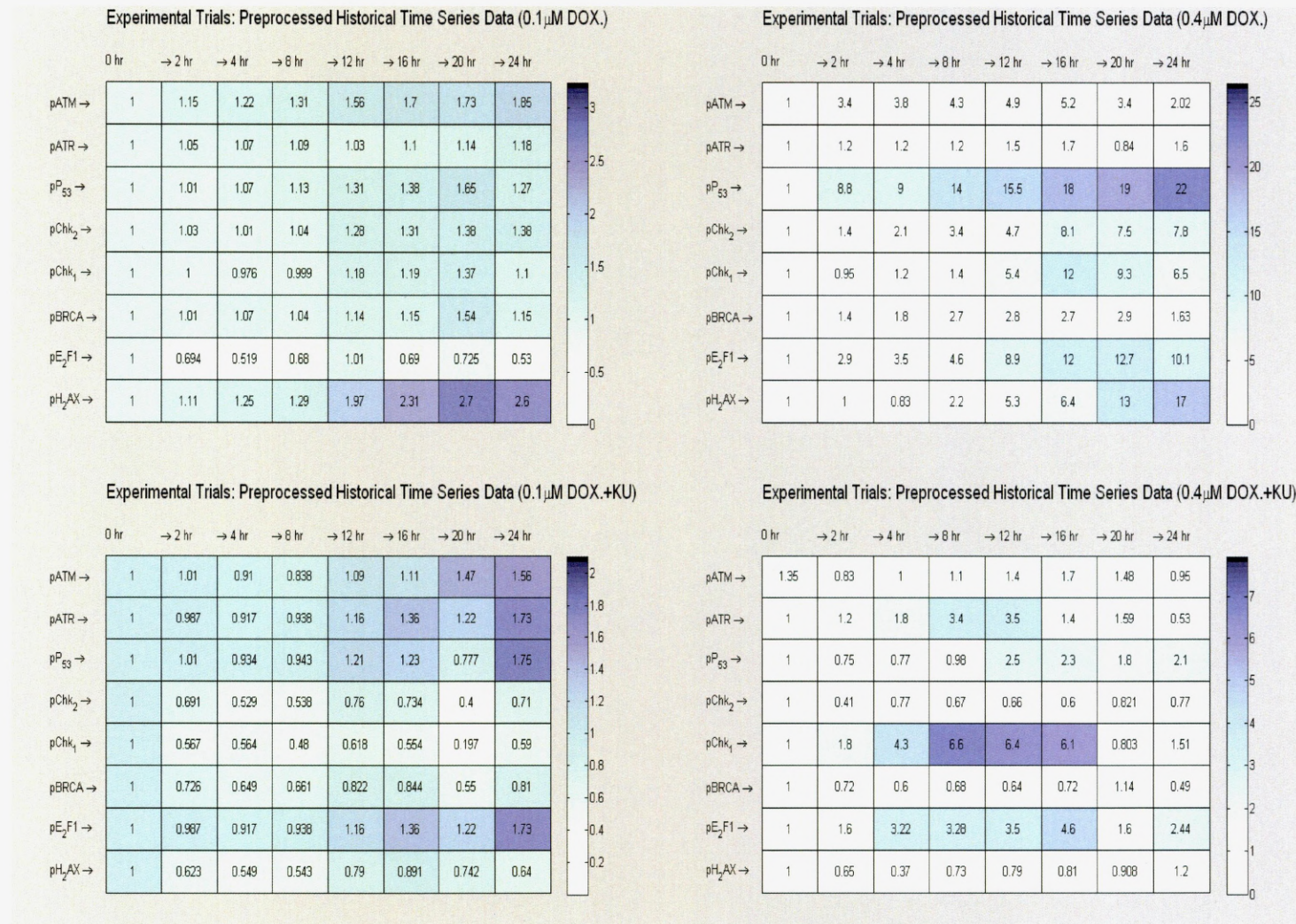


Figure 5-2: Time series measurements at lower (0.1 μ M) and higher (0.4 μ M) dose-intensities of both doxorubicin and doxorubicin+KU treatments.

5.3.4 Application of modelling methods

To understand the mechanisms by which ATM/ATR may be controlling cellular response to DOX input with and without ATM inhibition I analysed time measurements of pATM, pATR, pP53, pChk1, pChk2, pBRCA, pE2F1, and pH2AX, recorded at timepoints 0, 4, 8, 12, 16, 20, and 24 hours. Instant analysis (i.e. heatmap visual representation) of the time series data may be performed such as the heat map representation of values of key DNA damage response protein kinases (e.g. pATM, pATR, pP53, pChk1, pChk2, pBRCA, pE2F1, and pH2AX) displayed below in figure 5-2. The first task is to infer from each data set a predictive and data-consistent model that simulates exactly the given data in an unsupervised fashion. This predictive model may be constructed from the data by an appropriate reverse engineering method such as the transposive and repressive regression methods introduced in chapter 4 (Idowu M.A. 2011*b*, Idowu M.A. 2012).

The dynamical models are constructed from experimental data of 7 timepoints only, i.e. 0, 4, 8, 12, 16, 20 and 24 hours. First we ensure that every constructed model is data-consistent by ensuring that the data simulated from each of the constructed predictive models is compared with the original experimental data. An initial assessment of the constructed model is performed to see if the simulated and original data match. Models that do not pass this initial assessment test are then discarded. We applied the transposive regression method (Idowu M.A. 2011*b*, Idowu M.A. 2012) to estimate model parameters and ensure data consistency.

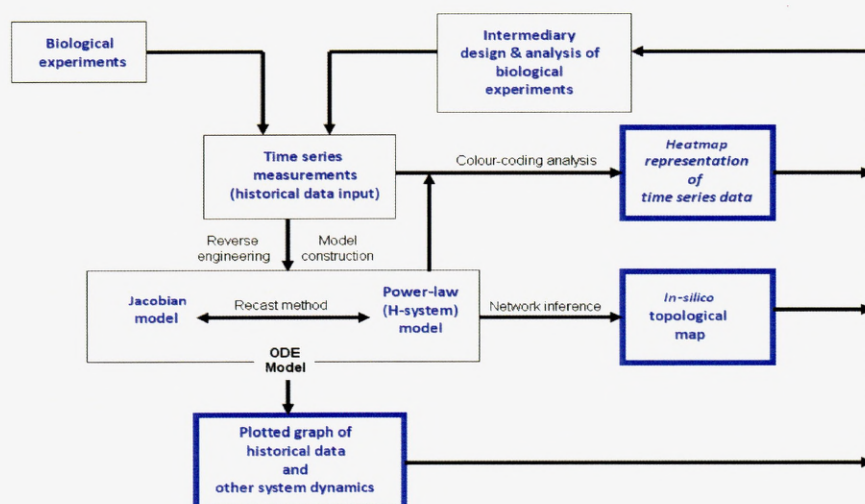
Once the constructed model is produced, inferring the underlying network of interaction from such a network model is routine. The major challenge encountered is in the area of system identification and parameter estimation.

Visualisation of network of interactions (for examples network diagrams in Figures 5-9 to 5-12) illustrated in Figure 5-3 is also required to carry out topological analysis in order to gain understanding and new insights into the biological pathway. This analysis of the modelling results may help establish gain new understanding and gain insight into the key underlying processes behind DNA damage response.

We use Graphviz (Software 2011), an open source graph visualization software, to generate the topological map representations of the constructed network models. The descriptions of the graphs are specified in simple (raw) text files that the Graphviz layout programs can read to draw network diagrams either in Postscript or Pdf formats. The Graphviz visualisation software does not provide a GUI editor but it has a number of layout tools for making aesthetically pleasing drawings depending on what the user needs. The default tool for drawing directed graphs is the “Dot” tool which uses the Dot language. We have selected the Dot tool and applied to visualise all the network data used in all the case studies presented in this thesis.¹

5.3.5 Use of heatmaps

Instant characterisation of the dynamics of biological systems may be represented by heat maps of table of time series data. In such heatmaps individual values are represented with shades of colours coded in magnitudes and depicted with a range of light and dark colors, representing low and high values, respectively. Such instant heat maps may be used across a number of comparable time series datasets representing cells in different states.



¹We do accept the fact that some of the network diagrams require some readjustments in the way the maps have been drawn. Therefore we recommend that a lot could still be done to make the drawings aesthetically pleasing.

Figure 5-3: A new dynamic modelling and reverse engineering strategy for analysing time series data of DNA damage response pathway to infer and construct a data-consistent predictive model of the system.

5.4 Analysis and interpretation of results

This framework establishes a modelling infrastructure for analysing experimental data that accounts for the interactions among the measurables over time, represented as a single interaction network with weighted strengths. Context-sensitive interpretation, i.e. by the (biological) domain expert, of the patterns in the network structure is then possible. Networks might reflect feedforward and also feedback. Interrelationships between the different signalling nodes are represented by links (directional arrows) and each (parameter) is assigned a numerical value to depict the strength of the interactions; the stronger the interaction the larger the magnitude of the parameter value specified.

The following subsections contain information, such (extractable) information as inferred from data, for the visualisation of *in-silico* model predictions. The topological maps extracted from the inferred data-consistent network models are shown to approximate the biological system. These maps represent the interactions between ATM and ATR and their immediate downstream substrate pP53 and other subsequent substrates, e.g. pChk1, pChk2, E2F1, BRCA1, and H2aX. To aid interpretation, each of these signalling proteins is categorised into one of the following layers: double stranded breaks, DNA, tumour suppressor, cell cycle arrest, apoptosis, DNA repair (gene mobilisation and repair), and transcription factor.

5.4.1 The constructed jacobian and Half-system models

Figures 5-4 depict the set of jacobian matrices representing the inferred jacobian models constructed using the transposive regression algorithm created in chapter 4. The matrices in figure 5-4 are inferred straight from the given time series data. Figures 5-4 represents the constructed jacobian models of the four systems 0.1 μ M Dox (top-

left), 0.4 μ M Dox (top-right), 0.1 μ M Dox+KU (bottom-left), and 0.4 μ M Dox+KU (bottom-left).

5.4.2 Initial analysis of data

As expected, a relative rise in levels of both pATM and pP53 was observed at 0.4 μ M and 0.1 μ M Dox, i.e. without ATM inhibition. ATM inhibition was more pronounced at 0.1 μ M Dox and early point of 0.4 μ M Dox input than at any other times. This suggests that ATM switches roles at lower dosage or early time points of higher dosage with Dox. However, upon higher dosage (0.4 μ M) with Dox, ATM may then switch from cytoprotective to cytotoxic role when inhibited. These observations confirm what was originally proposed by the domain expert that ATM may be playing both cytoprotective and cytotoxic roles. In addition to this confirmation, dynamic modelling may be used to determine the time-period at which the switch from cytoprotective to cytotoxic function may be happening.

Figures 5-5 to 5-8 show the relation between the experimental data (points) and simulated data generated from data consistent model (lines) for each of the proteins measured. As shown, the model - continuous in time - fits well to the discrete experimental data points and so is a data-consistent representation of the data.

Figure 5-7 was the most challenging analysis because of the inapparent co-dependency between the values of ATR and E2F1. Model predictions were most difficult in this case.

Since both the figures 5-7 and 5-8 show oscillatory patterns in dynamics, it may seem inappropriate that time invariant models be used to describe such systems. As suggested by the domain expert, such biological data should be split smaller segments (e.g. 0-8hr, 8-24hr, etc) to enable further analyses at different time periods to uncover the underlying signalling relationships in more details.

Figures 5-9 to 5-12 are derived from the topological data displayed in figure 5-4, i.e. each topological map is a network representation of the system represented by a transformation matrix specified in figure 5-4. There is a direct (1-to-1) connection between these topological maps and those four dynamic models inferred from the

experimental data.

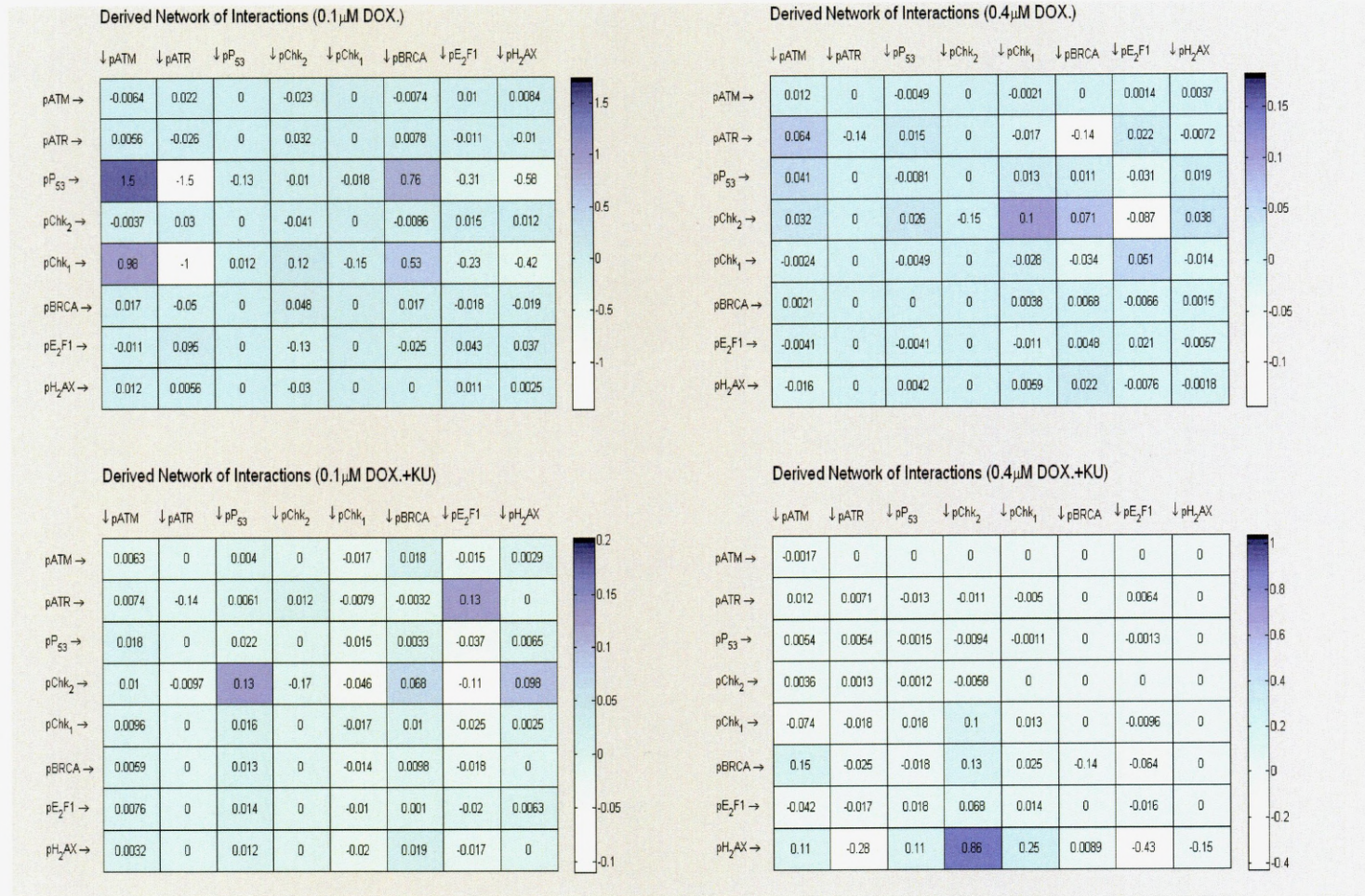


Figure 5-4: The reverse engineered jacobian models that are consistent with historical time series measurements at lower (0.1μM) and higher (0.4μM) dose-intensities of doxorubicin with and without KU treatment.

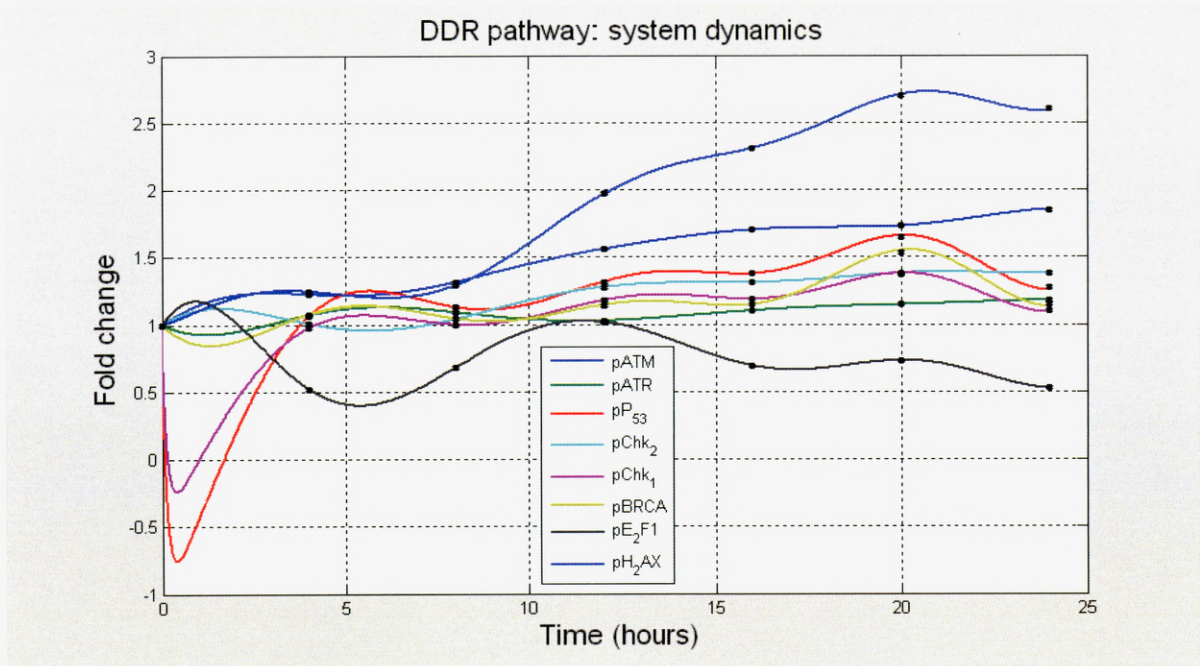


Figure 5-5: Simulation of system dynamics: consistent with historical time series measurements at (0.1 μ M) dose-intensities of doxorubicin without KU treatment.

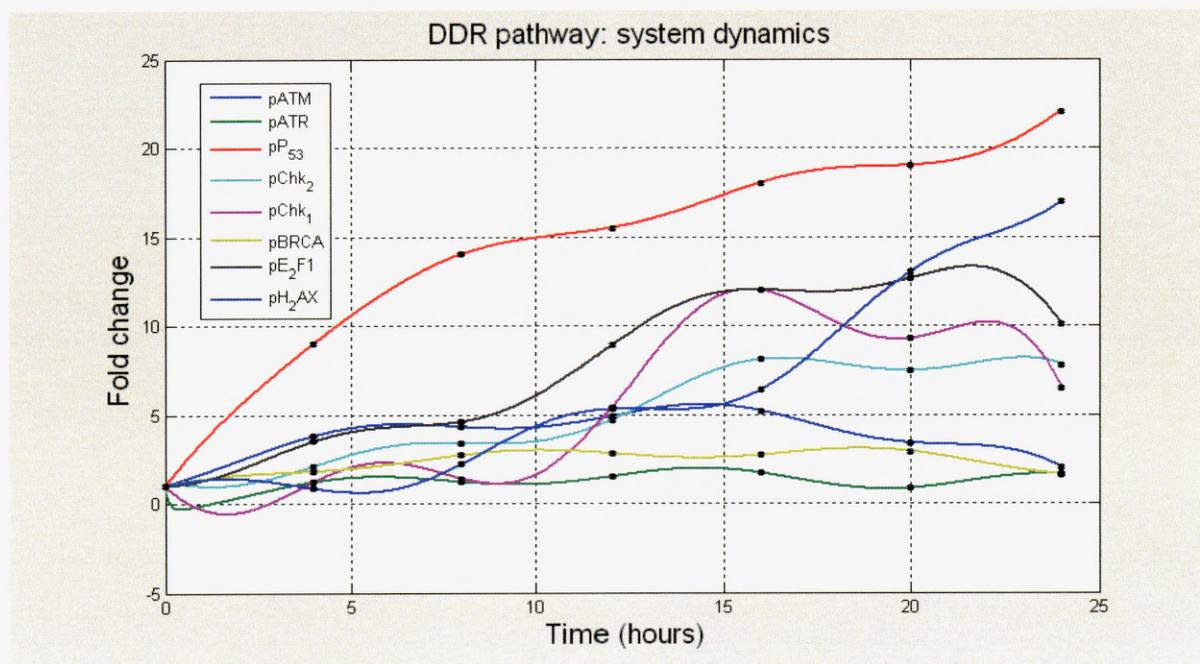


Figure 5-6: Simulation of system dynamics: consistent with historical time series measurements at (0.4 μ M) dose-intensities of doxorubicin without KU treatment.

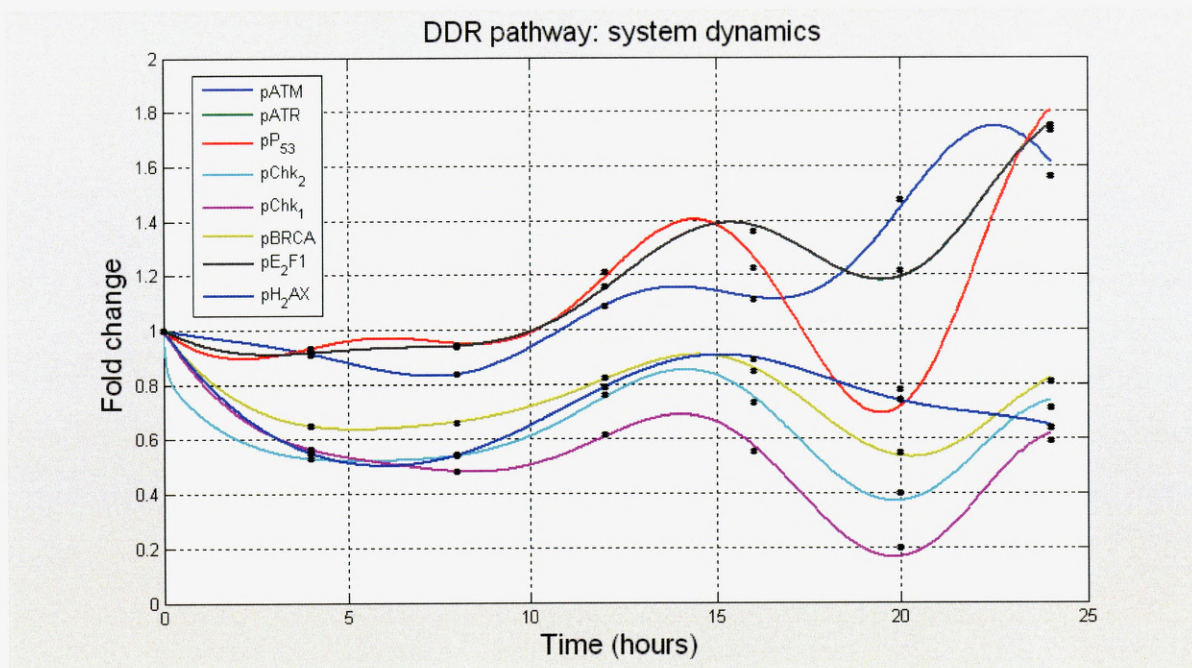


Figure 5-7: Simulation of system dynamics: almost consistent with historical time series measurements at (0.1 μ M) dose-intensities of doxorubicin with KU treatment.

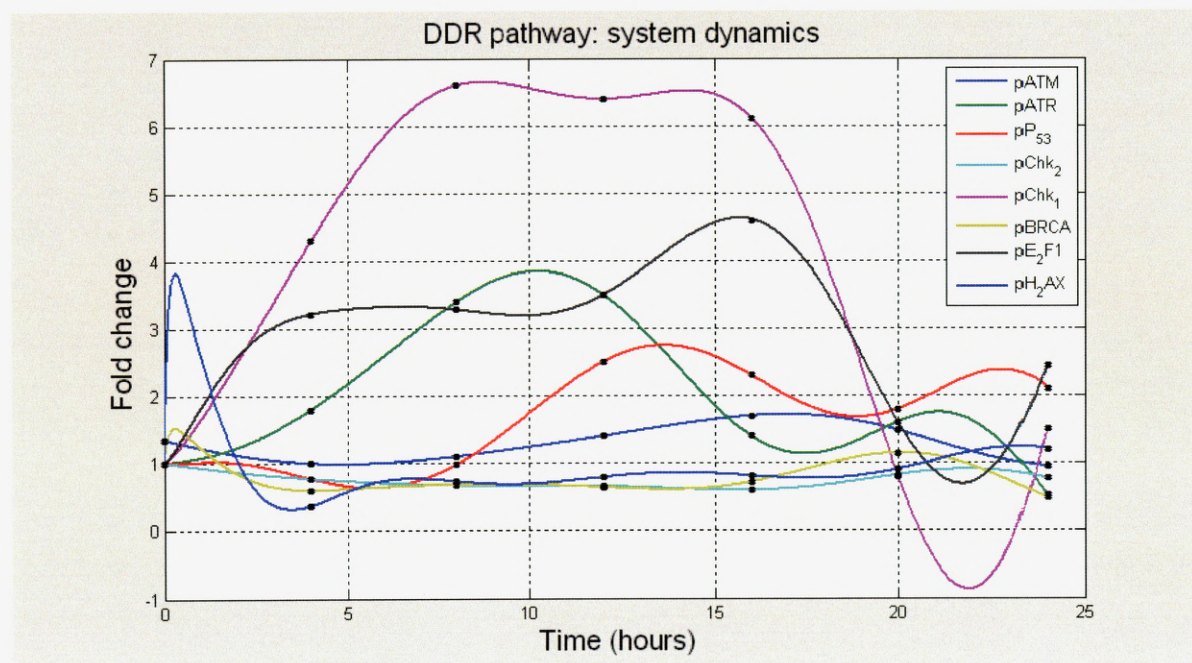


Figure 5-8: Simulation of system dynamics: consistent with historical time series measurements at (0.4 μ M) dose-intensities of doxorubicin with KU treatment.

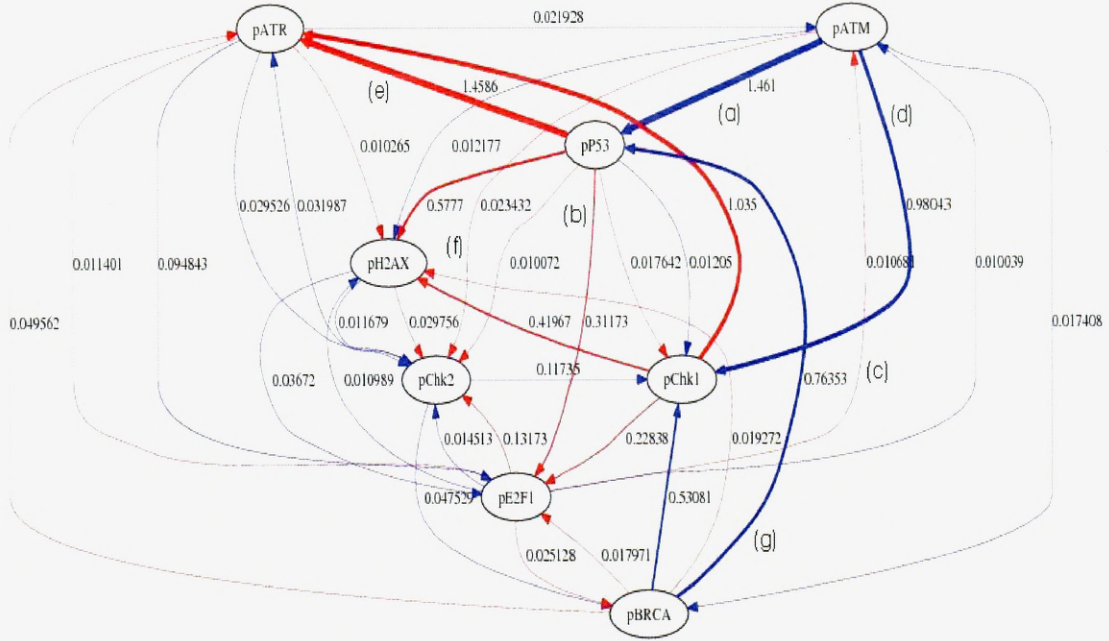


Figure 5-9: Derived topological map of network of DDR signalling pathway at 0.1 μ M Dox input in the absence of ATM inhibition.

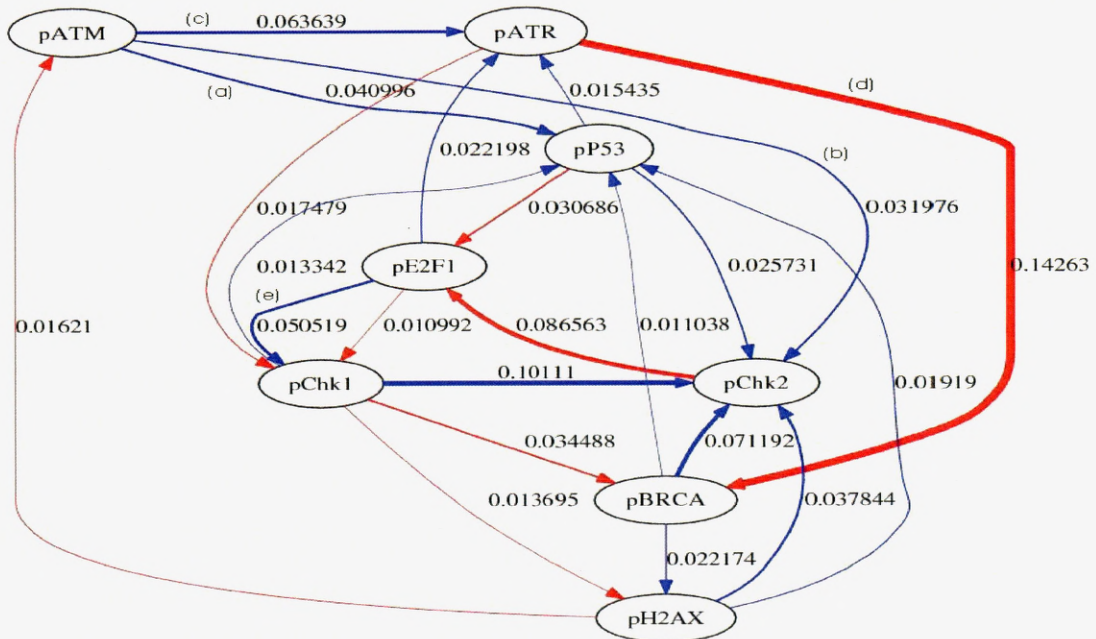


Figure 5-10: Derived topological map of network of DDR signalling pathway at 0.4 μ M Dox input in the absence of ATM inhibition.

5.4.3 Interpretation of results

Figure 5-9 represents the derived topological map of network of DDR signalling pathway at $0.1\mu\text{M}$ Dox input in the absence of ATM inhibition. In the absence of ATM inhibition and at $0.1\mu\text{M}$ Dox input (DNA-damage) the edges (links) depict various interactions that can be explained biologically. (a) ATM activity is induced by that level of damage with P53 shown to be an immediate substrate of ATM (Banin et al., 1998) (b) downregulation of E2F1 by the induced pP53 which may be interpreted to mean pP53 induction by ATM results in cell cycle (G1/S) arrest and sequestration of E2F1 by Rb (Chehab et al., 2000, Maya et al., 2001) possibly suggesting E2F1 inhibition mediated by pATM which may lead to cell cycle arrest as evidenced by the negative link between pATM and E2F1 (c) and pATM-mediated induction of pChk1 (Ho et al., 2004). The figure also shows pATR inhibition (e) by pP53 as reported in previous studies in literature, e.g. downregulation of ATM expression by P53 (Claig et al., 2010), transcriptional upregulation of Cyclin G by P53 which results in PP2A recruitment (Okamoto et al., 1996) that eventually suppresses ATR activity (Leung-Pineda et al., 2006, Petersen et al., 2011), upregulation of WIP1 by pP53 (Fiscella M et al., 1997) which reverses the ATR-mediated DDR pathway (Lu X et al., 2005). Such pP53-mediated upregulation of WIP1 may also contribute to the suppression of pChk2 (Fujimoto et al., 2005) and -H2AX (Moon et al., 2010) as shown in (f). This result also indicates that ATM signalling at this lower scale of damage may be characterised by both pBRCA1-mediated (g) and ATM-mediate induction of pP53 (a) -leading to cell cycle and DNA repair. Overall, these results indicate that at $0.1\mu\text{M}$ Dox level of DNA damage the cell may avoid triggering apoptosis by regulating the activities of pChk1, pChk2 and pATR and cell cycle arrest and DNA repair. Finally, these results indicate ATM's roles in cytostatic and cytoprotection against genotoxicity to promote DNA repair at a lower scale of DNA-damage (Khalil 2012).

As commented by Dr Khalil, since the cytotoxicity assay showed lower cell death at this point as compared to cells with a blocked ATM function, this up regulated pP53 may exert CIP/KIP mediated cell cycle arrest and promote ATM dependent repair

of the damaged DNA.

Figure 5-10 represents the derived topological map of network of DDR signalling pathway at $0.4\mu\text{M}$ Dox input in the absence of ATM inhibition. Notice that the positive signal between ATM and P53 (a) is disrupted in figure 5-12 (no link between ATM and pP53) where ATM is inhibited at this same scale of DNA damage. A link (b) that represents the induction of pChk2 by ATM, which does not exist in figures 5-9 and 5-12, may indicate Chk2-dependent apoptotic signalling (Hirao et al., 2002, Rogoff et al., 2004). The presence of ATM-dependent (c) ATR-mediated (Adams KE et al., 2006, Jazayeri A et al., 2006) suppression or inhibition of BRCA1 (d) possibly may be contributing to the greater apoptotic signalling observed after 12 hours, because ATR-mediated phosphorylation of BRCA1 promotes cell cycle and DNA repair (Tibbetts et al., 2000, Zhu et al., 2006). As Figure 5-12 indicates absence of ATR-mediated inhibition of BRCA1, it might be worth investigating if such system perturbation could be responsible for the lower apoptotic signalling that is evidenced in ATM inhibited state between the same 12-24hr period. Another unique signal observed here and that does not exist in figures 5-9 and 5-12 is the link (e) that suggests indirect activation of pChk2 by E2F1 via induced pChk1. This may be a key difference in this figure and figure 5-12. Another remarkable difference between this figure and figure 5-9 is that here apoptotic signalling may be triggered by ATM signal to pChk2 (instead of ATM signal to pChk1 which causes cell cycle arrest) (see link (b)). pChk1 is shown to positively signal to pChk2 here and in figure 5-12.

Dr Khalil comments that ATM inhibition at a higher scale of damage (between 12-24hr treatment) showed lower cell death as compared to when ATM was functional. Interestingly, while at lower scale of damage in ATM inhibited state, the new positive link between E2F1 and pATR was proposed to be promoting cell death, at a higher scale of damage in ATM inhibited state, E2F1 levels were shown to have a repressive effects on ATR. Theoretically, this supports the earlier presumption and suggests that the lower cell death seen during ATM inhibition at higher scale of DNA damage may be the absence of E2F1 induction of ATR.

Figure 5-11 represents the derived topological map of network of DDR signalling

pathway with combinatorial treatment at 0.1 μ M Dox and with ATM inhibitor. The ATM-induction of pP53 appears less strong (a) when ATM is inhibited than without (figure 5-9). The indicated inhibitory influences that pP53 exerted on pChk1, pChk2, E2F1 and γ -H2AX in the absence of ATM inhibition now appear to be reversed to positive in ATM inhibited states. The experimental data and dynamics of the system seem to suggest a later induction of pATR and E2F1 in response to ATM observation. Also a positive signal from E2F1 to pATR (e) is indicated which is associated with greater apoptotic activity commonly attributed to states with ATM inhibition. Alternatively, the disruption of ATM and loss of pP53-mediated inhibitory influence on pATR by pATM (previously shown in figure 5-9) may lead to further apoptosis. Furthermore, this result also suggests that pP53-mediated suppression of E2F1 by pATM (d), which can lead to cell cycle arrest, may eventually cease E2F1 sequestration to upregulate ATR (e) that further triggers apoptosis when ATM is disrupted, an outcome that is consistent with the fact that during 0.1M Dox alone (as indicated in figure 5-9) downregulation of pATR by pP53 and BRCA1 is thought to imply suppressed apoptotic activity and enhanced DNA repair. ATR activity has been reported to signal apoptosis before (Kumar et, al., 2005, Pabla et al., 2008). These results indicate that DNA-damage at this scale of damage may result in sequestration of E2F1 which could lead to pATR upregulation (e). Hence this conclusion is consistent with the previous explanation about how ATR inhibition by pP53 (as indicated in figure 5-9) might result and lead to DNA repair at lower scale of damage.

Dr Khalil remarks, these experiments have not only revealed the effect of ATM kinase inhibition on cellular sensitivity to time course treatment of Dox, but also uncovered the underlying signalling network that is suggested to promote ATM mediate cell cycle arrest and DNA repair at a lower scale of DNA damage, and the concomitant signalling alterations and appearance of novel links following ATM inhibition that were suggested to influenced the observed alteration in the cellular sensitivity process. These results further indicate that greater sensitivity via DDR manipulation brought about by ATM kinase inhibition is caused by sequestration of E2F1 which

otherwise would lead to upregulation of pATR activity.

Figure 5-12 represents the derived topological map of network of DDR signalling pathway at $0.4\mu\text{M}$ Dox input in the presence of ATM inhibition. The presumed positive signal between ATM and P53 in figure 5-10 is totally disrupted here. This figure indicates the absence of ATM-mediated induction of pChk2 and absence of ATR-mediated inhibition of BRCA1 both of which are present in figure 5-10. Not only that, there exists no indirect activation of pChk2 by E2F1 via pChk1 here. pChk1 is shown to slightly positively signal to pChk2 here and in figure 5-10. BRCA1 is shown to negatively signal to E2F1 thereby lowering apoptotic signalling (a).

5.4.4 Further analyses of segments of experimental data

In-silico topological maps may be determined based on full time series data sets supplied or their partition sets that have been divided into non-empty (non-overlapping) subsets. Sometimes experimental designs or oscillations observed in system dynamics might require that transient topological maps be captured at partition intervals to express the transient behaviours and multiphasic signatures in biological signalling networks. As agreed with the domain expert, the recommended methodological approach to dynamic data that reveal complex oscillatory patterns (i.e. figure 5-8) should be simple: split the biological data into the desired dimensions of non-overlapping data-segments (e.g. 0-8hr time point and from 8-24hr time point) to enable further capture of intermediate transitivity in dynamics and then reapply the developed inference method to delineate biphasic ATM function and uncover the underlying signalling relationships and the associated alterations during the shift of cellular response.

The analyses were performed as suggested and the results are then displayed in figures 5-13 and 5-14. In those figures, only the $0.4\mu\text{M}$ Dox input information (on the right hand side) which contain the inferred models of DDR signalling pathway at $0.4\mu\text{M}$ Dox input in the absence and presence of ATM inhibition were analysed because those were the data sets that exhibited pronounced oscillations in their dynamics. Their derived topological maps are displayed in figures 5-13 and 5-14.

Figure 5-13 represents a further derived topological map of the network of DDR

signalling pathway at $0.4\mu\text{M}$ Dox input in the presence of ATM inhibition using a subset of data with time points 0hr, 2hr, 4hr, and 8hr. ATM functions in a cytotoxic role.

Figure 5-14 represents further derived topological map of network of DDR signalling pathway at $0.4\mu\text{M}$ Dox input in the presence of ATM inhibition using a subset of data with time points 12-24hr. ATM functions in a cytoprotective role.

Figures 5-13 and 5-14 represent the derived topological maps of network of DDR signalling pathway at $0.4\mu\text{M}$ Dox input in the presence of ATM inhibition. Specifically, inhibition of ATM resulted in higher apoptotic cell death only at 2, 4 and 8hr of Dox treatments as compared to cells treated with Dox alone. Extension of treatment to 12, 16, 20 and 24hr switched the outcome of ATM inhibition where lower cell death was seen in Dox treatment with ATM inhibition than without. Apoptotic signalling is noticed at 2, 4 and 8hr of this treatment (figure 5-13). A switch to lower apoptosis is then observed after 8 hours as compared (figure 5-14) to treatment without ATM inhibition.

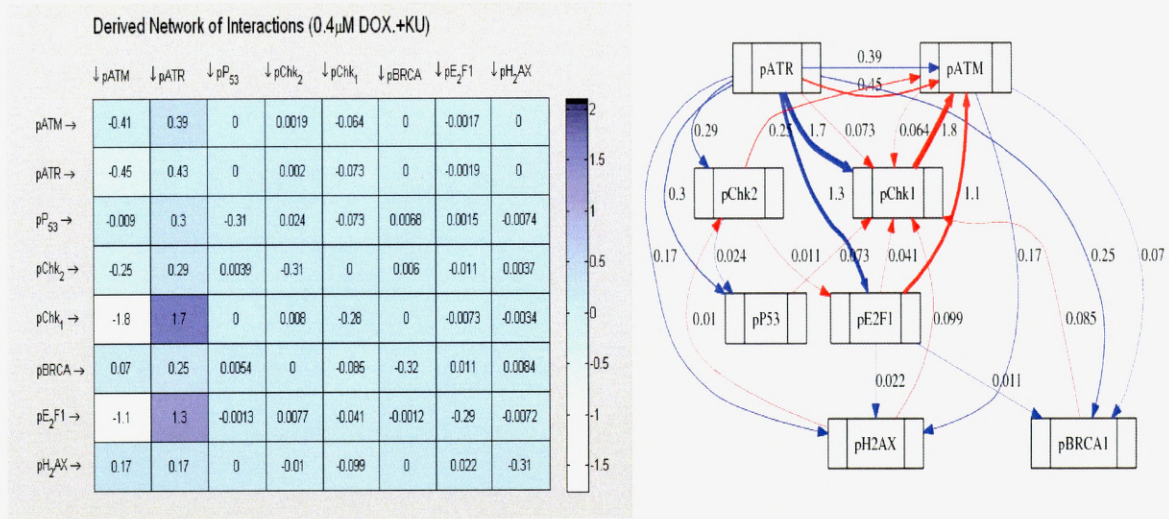


Figure 5-13: Further analysis: reverse engineered jacobian models that are consistent with historical time series measurements at (0.4 μ M) dose-intensities of doxorubicin with and without KU treatment using data with timepoints 0-8hr.

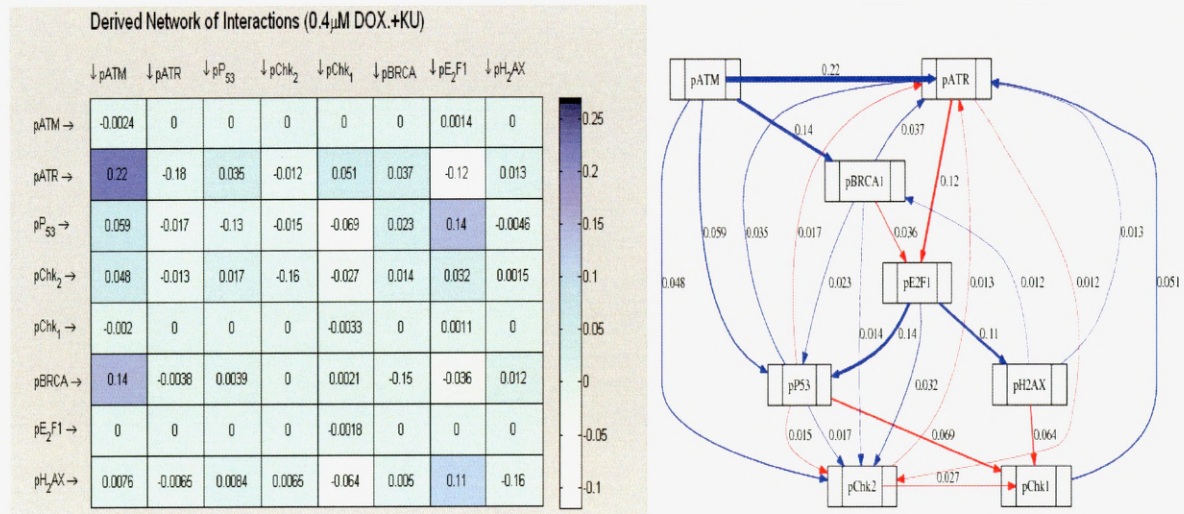


Figure 5-14: Further analysis: reverse engineered jacobian models that are consistent with historical time series measurements at (0.4 μ M) dose-intensities of doxorubicin with and without KU treatment using data with timepoints 12-24hr.

5.5 Discussion and conclusion

We investigated the DNA-damage response pathway involving a number of signalling proteins (pATM, pATR, pP53, pChk1, pChk2, pBRCA, pE2F1, and pH2AX) under 0.1 μ M and 0.4 μ M of Doxorubicin, with and without ATM inhibition by KU. Relative quantification of individual proteins were measured after 2 hours and regular time intervals of 4 hours with the individual data based on the initial condition. The modelling approach was completely data-driven and entirely based on analysis of the time series profiles of the proteins involved in the DDR pathway. No *a priori* knowledge about was given prior to the system identification process. The results themselves suggested some new insights and confirmed known results thereby demonstrating and validating the effectiveness of our data-driven modelling approach. The results we obtained are summarised below. Those new insights were obtained from comparative study of the topological information derived from structures of multiple network models inferred from experimental data. We inferred optimal data-consistent jacobian models of the system using the method of the transposive regression TRM. Then topological maps of interaction network were inferred from those models. 2D images of those maps were analysed further to determine cellular responses to intervention under inhibitory and non-inhibitory conditions of ATM.

The following key signalling determinants are inferred purely from experimental data. We infer that ATM upregulates pChk1 and BRCA1 at the lower scale of Dox treatment causing cell cycle arrest and repair in ATM non-inhibited state and ATR positively influence pChk2 in ATM inhibited state. With E2F1 levels induced, E2F1 may be activating pChk2 indirectly (via pChk1) resulting in greater cell death. Disruption in ATM activation links to p53 at a lower scale of Dox treatment with KU and may lead to stronger ATR upregulation and eventually apoptotic signalling. ATM inhibition resulted in lower cell death as compared to ATM active state at higher scale of Dox treatment suggesting that ATM may be playing a cytotoxic role at higher Dox treatment.

The *in-silico* prediction of system dynamics based on data of experimentally de-

terminated fold induction of DNA damage response pathway also revealed oscillatory patterns in the modelling results of the system in non-inhibited ATM state. These modelling results may be suggesting the involvement of ATM kinase activity in downstream protein oscillations following double stranded DNA damage [according to model result interpretation by the domain expert]. According to the domain expert Dr Khalil, oscillation of components of intra-cellular signal transduction pathways is a feature of feedback loops necessary to maintain system equilibrium. For example, in previous biological experiments and computational modelling studies involving oscillatory phenomena in proteins (e.g. protein activation via recurrent initiation mechanism associated with feedback loops) at different concentration of genotoxic agent had been reported before [Lev Bar-Or et al., 2000, Lahav, 2004, Batchelor et al., 2008, Ma et al., 2005] as well as associated physiological consequences, particularly those that had direct impact and exhibited significant influences on cell mortality in response to DNA damage (Sun et al., 2009, Zhange et al., 2009). Therefore, the modelling results of experimental data, as supported with evidence in literature [Lahav et al., 2004, Xhang et al., 2009], suggests induced disruption of feedback loops at both lower and higher levels of DNA damage when ATM is inhibited.

The domain expert Dr Khalil also remarks, ATM functions both in cell cycle arrest and DNA repair whereby it promotes cytoprotection (Bao et al., 2001, Lim et al., 2001), as well as functions as a central component in triggering apoptotic cell death (Westphal et al., 1997, Chong et al., 2000, Powers et al., 2004) and its specific downstream signalling preference is context dependent, ATM may have a cytoprotective effect during Dox time course treatment up until the 8hr time point (as cells were sensitised to ATM inhibition) and may switch its function and downstream signalling preference to apoptotic mode when the damage is enhanced at post 8hr time points (as lower cell death was seen during ATM inhibition). Network inference results confirm cytoprotective role of ATM in damage repair and survival at lower scale of DNA damage. ATM inhibition resulted in lower cell death as compared to ATM active state at higher scale of DNA damage. As expected, models of ATM pathway show different signals in different treatment therapies and data segments (i.e with

different time segments). Additional time point data may be required to improve prediction results and explain the key signalling patterns and potential molecular targets that should be considered in line with the appropriate dosages of Dox and/or KU treatments.

Clearly effective system identification and parameter estimation methods such as the inference method demonstrated here are useful tools that may help understand various key factors, mechanisms and dynamics of the induced DNA damage response pathway that may be responsible for the cell to survive. Such methods are useful and may be applicable to the design and inexpensive assessment of potential therapeutic treatments, e.g. elucidating the role of ATM in damage response pathway more clearly after Dox treatment. As demonstrated in this report using a worked example, there is strong evidence to support the effectiveness of the modelling approach used in analysing real life experimental data.

Chapter 6

Dynamic modelling of PI3K-AKT signalling pathways

Over the years diverse computational and mathematical methods for modelling biological systems have been developed. The idea of using time series experimental data to model and understand signal transduction pathways seems attractive. This is due to the advancement in proteomics technologies. Today concentrations can be measured relatively easily and recorded for modelling purposes. Likewise if the modelling objective is to consider mRNAs at the genetic level, the experimental data for such biological research can be acquired using microarrays technologies - the latest advances in microarray technology make it possible to obtain large volumes of data in ways never imagined before. In systems biology modelling of signal transduction pathways continues to improve with numerous contributions from researchers all over the different fields involved.

Cancer systems biology involves both the understanding of causes and nature of cancer, the development of new technologies, and application of systems approach to cancer treatments. Though it involves the integration of multiple biological scales (e.g. molecular (signalling), cellular, tissue, organ, system, organism), this holistic approach is not commonly practised due to the complexity and enormous challenges involved in the complex interactions within and between cells, tissues, and organs. The fact that cancer studies may provide useful answers to questions asked at a

molecular level may not mean challenges faced at other biological levels above it are automatically solved. Therefore, it is fitting to think about how to easily translate solutions to difficult challenges faced at a molecular level to the other biological levels above it. Adequate research time may be required to carry out research at each of the biological scales already specified. The modelling challenges encountered at each of these levels are in themselves extremely difficult barely leaving room to adequately address other important issues that relate to scalability. However, key processes that relate to signalling, apoptosis, DNA repair, cell cycle, cell growth and survival, must be addressed first at both the molecular and cellular levels. It is the understanding of such key processes, the various multivariate dysregulation involved in cancer formation, and prospective treatments available that is vital in cancer systems biology (Kreeger P.K. 2010). Because cancer is highly complex and heterogenous in nature with diverse genetic mutations and multiple dysregulated pathways, researchers seek to use mathematical modelling to understand how current understanding of cancer causes, development, and treatments may be improved.

Some of the challenges confronting cancer systems biology include inadequate or lack of understanding about the impact and consequences of critical molecular alteration on tumour cell phenotype, different tissue types, and diverse multiple effects on organs in response to systemic changes (Kreeger P.K. 2010). At the molecular and cellular level, a number of extracellular and intracellular signalling pathways may be dysregulated. One such system is the ErbB/HER signalling system which is frequently mutated in cancer and has been the subject of numerous studies (Kreeger P.K. 2010, Citri A. 2006, B.S 2005). In this case study we seek to investigate and understand the ErbB/HER-PI3K-MAPK signalling pathways.

In addressing important issues such as targeted drug therapies, many would agree mathematical modelling has great potential in supporting drug design. However, mathematical modelling is yet to adequately and efficiently deliver this potential, generally and specifically for drugs that target the human epidermal growth factor receptor (HER) signalling pathway in the treatment of cancers, especially those that are characterised by HER2 overexpression. The HER family, being an important and

frequently studied signalling network (Oda K. 2005, Citri A. 2006, Soltoff S.P. 1994, Kim H.H. 1994), comprises members that are often implicated in human cancers. For example, the HER1, HER2 and HER3 receptors are commonly overexpressed in cancers (Slamon D.J. 1987, Slamon D.J. 1989, Naidu R. 1998, Stephens P. 2004) and this has motivated researchers to seek to employ mathematical modelling to study their activation. Such studies focus on the tyrosine kinases and subsequent molecular targets downstream of the signalling cascade such as PI3K and MAPK pathways. The goal of such studies would be to determine how best to develop a therapeutic strategy aimed at a cancer treatment objective, e.g signalling of apoptosis in only cancer cells, restoring aberrant functioning to normal functioning, etc.

Most of the computational and modelling methods that target the HER, PI3K, and MAPK signalling pathways are process-based. For example, comprehensive details about the underlying dynamical system may be formulated to construct an ordinary differential equation (ODE) model of the biological system. Such predictive models require that all key proteins have the necessary specification and formulation expressed within the appropriate compartments of the model and can be useful in simulating semi real data. The data produced from such surrogate experimental systems, which may be a little different from real experimental data but very similar in many ways in the sense that the processes that have been captured and embedded in them very much resemble those in the real systems, may be used to further test our modelling approach. Using plausible data such as those generated from a published process-based model (Goltsov A. & Harrison 2011) we further explore the capabilities of our inference algorithm, primarily to both demonstrate its value and explore some of its hidden weaknesses in a controlled environment.

Our aims in this case study are:

1. to acquire samples of time series data of HER 2/3-PI3K-MAPK signalling pathways simulated *in-silico* from a biologically plausible process-based system of a real system;
2. to automatically construct more simplistic alternative ODE models that are

able to capture and represent nonlinear system behaviours purely from the time-series data;

3. to ensure that each constructed model is data-consistent and can reproduce the exact data through simulation;
4. to use each constructed model to understand the HER2/3-PI3K-MAPK signalling pathways and effects of drug input like Pertuzumab (2C4), a monoclonal antibody for treating human cancers;
5. to use the models to identify the various molecular factors that may be contributing directly or indirectly to the development of sensitivity and resistance to cancer treatment with 2C4;
6. to offer recommendations on useful and novel cancer treatments that target the HER 2/3-PI3K-MAPK signalling pathways by formulating experimentally testable hypotheses that may answer any of the broad issues presented in figure 6-1.

In human pathology it may be necessary to use time series data of important biological pathways to decipher some of the very complex intracellular mechanisms that may have produced such data. To achieve this aim computational modelling can be employed. As indicated earlier in the previous chapters, the prediction of system dynamics from data can be a difficult and challenging task. Application of data-driven strategy is complementary to other modelling approaches and can be useful for the understanding of complex biological systems in which extracellular and intracellular signalling pathways are involved. Our focus is to demonstrate the importance and applicability of dynamic modelling (strictly based on data) rather than assumptions of processes in understanding intracellular signals emission and control within network of biological system using only simulated data of independent biologically plausible process-based models.

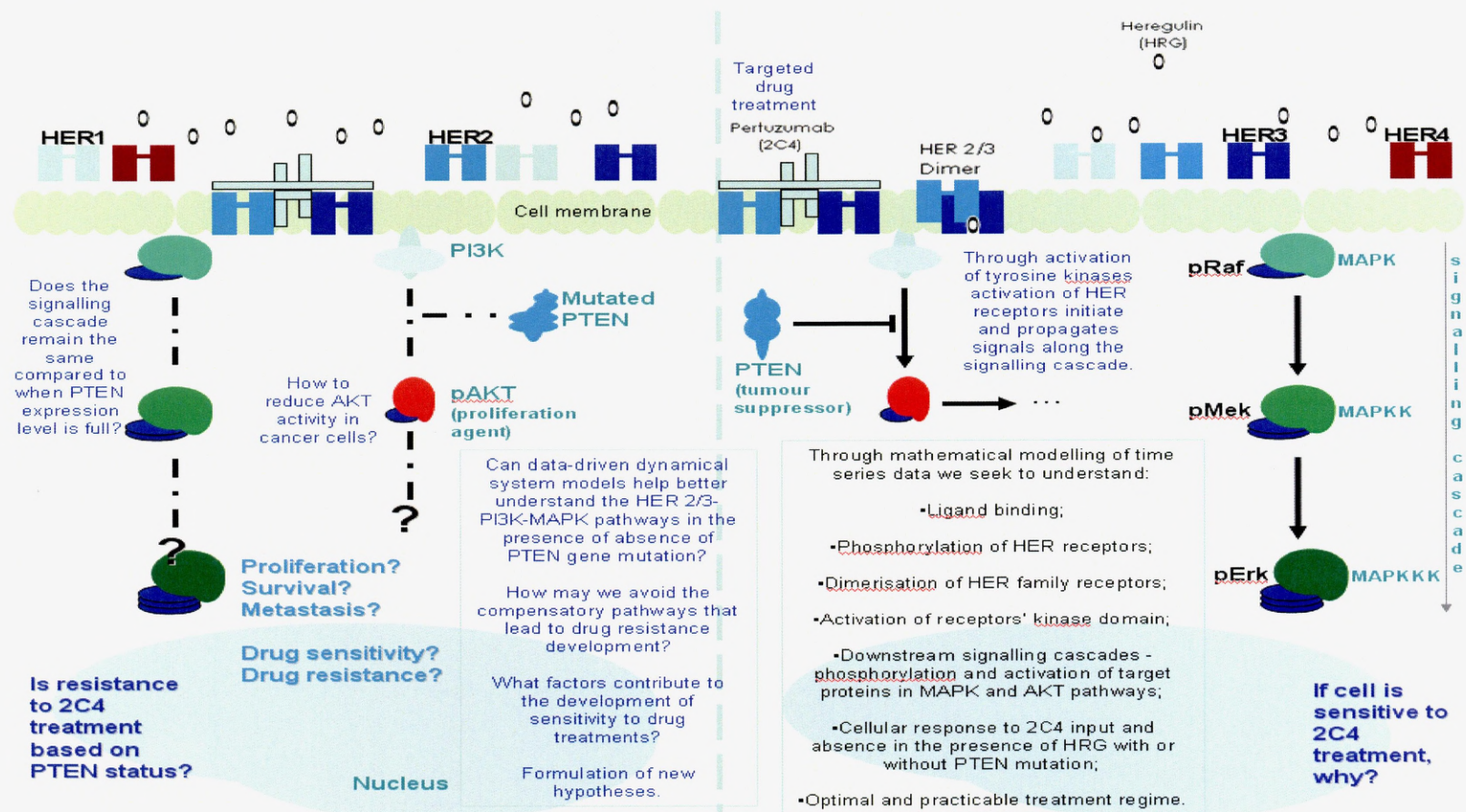


Figure 6-1: Schematic representation of the HER 2/3-PI3K-MAPK signalling pathways.

6.1 Background

In cells extracellular signal molecules may activate multiple intracellular signalling pathways at the same time. By so doing they may be said to control communication between and within cells. Located at the cell surface are special proteins, called receptors, that are responsible for both the reception and transmission of these signals to intracellular proteins (also called receptors). On contact and binding with extracellular molecules these cell-surface receptors are activated and initiate intracellular signalling pathways, while the intracellular receptors help make possible communication within intracellular pathways, i.e. intracellular signalling may trigger the activation of other protein targets e.g. effector proteins. These activation (or deactivation) of signalling pathways often occur and are constantly repeated across multiple pathways forming complex communication networks. As a result of these complex communication between and across multiple intracellular pathways it is difficult to understand how cell behaviour is being controlled and regulated. These signalling pathways involve metabolic pathway proteins and molecules and gene regulatory proteins.

Modelling of biological pathways may be initiated by describing the activities of the dysregulated pathways of a cellular system, characterising some of the most important key signal transduction processes of the complex system together with the various interactions among intracellular species (Kholodenko B.N. 2010, Kreeger P.K. 2009, Aldridge B.B. 2006). Often such systems have established known and unknown crosstalks (David Gilbert 2006), feedback loops (Papin J.A. 2005) within and across multiple pathways, alternative interconnectivities (Citri A. 2006) all forming highly complex topologies which are often robust to system perturbation during drug treatments (H 2002*b*, Goltsov A. & Harrison 2011).

Biochemical signalling (or absence of signalling) tends to manifest changes in the systems involved and these signalling systems are often viewed in terms of changes caused and effects propagated in them over a period of time (Alberts Bruce 2009). It is important to consider both the direct and indirect (alternative normal and abnormal) regulatory mechanisms by which the activity of specific or key proteins are

being controlled within or outwith the cells. Hence cancer modelling often starts by addressing the challenge of specifying or identifying aberrant signalling within and across a set of targeted pathways.

6.2 Understanding signal transduction

Through signal transduction the behaviour of the target cell is altered as cell-surface (or intracellular) receptors (acting as transducers) convert and transmit extracellular signals into intracellular communication. This communication primarily involves signalling activation of receptors and transduction of a network of other signalling proteins. (Alberts Bruce 2009)

Special signalling proteins may activate other signalling proteins either through phosphorylation (addition of a phosphate group to an organic molecule, e.g. protein, to alter its function and activity) causing many enzymes to be turned on and off. Others may be activated by dephosphorylation. To add to the complexity a single receptor may activate signalling pathways in parallel, and so may influence cell behaviour in a number of different ways (Alberts Bruce 2009). Understanding the various signalling mechanisms through which key changes may be effected is often a modelling focus.

For example, the Ras-MAPK and PI3K/AKT pathways are important pathways that control cell division, motility, and survival (Bown J. 2012). Because EGFR signalling always activates these pathways understanding the ErbB system and signalling within the Ras-MAPK and the PI3K/AKT pathways are common objectives of cancer treatment research. Network modelling aimed at experimental drug design may then focus on identifying aberrant signalling that may have contributed to the development of the cancer (Amit I. 2007) and also those that may propagate detrimental effects on cancer cell survival. Both of these objectives requires computational and mathematical modelling strategies. The main aims of modelling the ErbB system are to: describe input-output characteristics of the systems; determine the various key signalling responses to input signals and drug action; identify

the potentially active key targets for anticancer therapy; develop efficient strategies for identifying the mechanisms of drug sensitivity and resistance; and establish design and selection criteria for the optimisation of therapeutic treatment of cancer (Schoeberl B. 2009, Chen W.W. 2009, Faratian D. & Harrison 2009).

6.2.1 Major modelling challenges

Using mathematical and computational modelling to identify molecular targets from times series data is a difficult task, especially if required to be performed with a view to improve diagnosis and predict early prognosis of cancer. The need to adequately and correctly predict cellular responses to anticancer therapies demands that the systems biology modelling that is being used be able to identify and differentiate between the various highly complex topologies inherent in quantitative data acquired from cell lines which are either tumour-specific or normal-tissue related (Bown J. 2012). Our modelling objectives in this case study use purely time series data representing cellular systems and analyse them to infer and differentiate between different disease states. We therefore seek to predict and characterise cellular systems based on the molecular signatures of the signal transduction and associations among the system components. By exploring the applicability of the systems biology modelling strategy in this regards we may offer this solution as a complementary tool and approach to other forms of traditional modelling techniques. The complexity involved, apart from the inference of crosstalks across multiple pathways (David Gilbert 2006), the tasks of having to uniquely identify the various key feedback loops (Papin J.A. 2005), and alternate and compensatory pathways that help keep biological cells robust to system perturbation all complicate the modelling challenge.

We seek to capture the dynamics of the networks of the systems using most basic equations expressed using nonredundant parameters that uniquely and unambiguously describe the rates of change of concentrations of species involved in the network. We seek to offer insights into the performance of our new inference algorithms in terms of how their system identification abilities in differentiating between multiple alternate signalling networks revealed from data.

The mathematical modelling may be employed as a strategy for studying signalling systems' responses to stimuli. Such understanding of signal transduction mechanisms, (trivial and non-trivial) cross-talk communications across multiple signalling pathways, and signalling responses to drug interventions or perturbation in cellular systems may contribute to the development of effective strategies to inform the identification of therapeutic biomarkers (Goltsov A. & Harrison 2011).

From the computational and modelling viewpoint, the improvement in data simulation strategies, inference methods development and experimentation are gradually rising but not in line with rates of increase in the amount of data available. Among the list of recent approaches to modelling biological signalling network, the ordinary differential equations (ODE) approach is most attractive. However, it is important to note that modelling of signal transduction is challenging due to system identification and parameter estimation challenges. The notion that the modeller may decide upfront the entire topology can potentially invalidate system identification and parameter estimation if care is not taken.

6.2.2 Towards multiple parameter fits

Among a number of challenges the modeller has to face system identification and parameter estimation are the most difficult. The parameter estimation task can be difficult, time-consuming, and is often with no real success in a short time frame, especially if the model is process-based, i.e. requires a much deeper level of detail than simplistic models. In dealing with uncertainties sometimes it might be best to use a non iterative method that estimates multiple parameters all at once (Gutenkunst R.N. 2007). For example, Gutenkunst et al., in estimating the parameters of a growth-factor-signalling model, managed to narrow down the parameter search spectrum to a well-constrained domain. In a later work they reported each model examined in their collection had an unsystematic spectrum of parameter sensitivities that complicated the estimation process in predictive models. Their insights into the prevalence of sloppiness in parameter sensitivity spectra (Gutenkunst R.N. 2007) suggests that single-parameter estimation should be avoided and parallel parameter estimation methods should be

used whenever possible. They suggested that valuable time could be saved to focus on enhancing the predictive power of the model if the estimation strategy employed is appropriated. For this purpose we avoid single-parameter estimation whenever we can.

Our parameter estimation method offers more benefits than optimising single parameters one after another. Not only does it provide means of estimating all model parameters at once, it completely minimises the possibility of obtaining suboptimal results, thereby saving time and reducing model design or development cycle.

6.3 Reverse engineering of RTK-PI3K-MAPK signalling pathways

6.3.1 Problem definition

It is commonly known and accepted that the enzyme-coupled transmembrane receptor tyrosine kinases (RTKs) phosphorylate tyrosines on themselves and other intracellular signalling proteins (Alberts Bruce 2009). On binding to activated RTKs and activation, signalling proteins may send signals through multiple pathways which may then relay signals downstream (e.g. through Raf-Mek-Erk signalling) to the nucleus along the MAPK-pathway. The terminal kinase Erk may either phosphorylate and inactivate pRaf (thereby forming a negative feedback loop) or enter into the nucleus to phosphorylate gene regulatory proteins that may activate the transcription of other genes to effect changes in the cell. Alternatively other relay mechanisms for promoting cell survival and growth through activation of other pathways (e.g. PI3K) may be involved. However, such multilevel control of signal transduction from membrane receptors to nucleus may vary for different cell lines, depending on the interactions among the individual components of the signalling networks, whether the cells involved are cancerous or not, and in part on the expression level of the proteins involved (McCubreya J.A. 2006). Understanding the control and regulation of signal transduction and the mechanisms, development, loss of responsiveness (after treatment) and

acquired resistance to anticancer drugs may require novel computational and mathematical modelling techniques able to deal with the challenges associated with complex dynamical systems modelling. These challenges include inferring and identifying various alternative relay mechanisms for promoting cell survival and growth through PI3K activation, lack or mutation of PTEN in uninhibited PI3K signalling, cell survival and growth through the PI3K-AKT signalling pathways (Alberts Bruce 2009) and MAP kinase (Erk) signalling. The process-based model that generated the data used in this case study had been well-tested against real experiments and biological experiments and simulated results are in good agreement.

The problem definition is this: can our model inference algorithm differentiate between data obtained from models operating in “sensitive mode” and those in “resistant mode” in response to RTK inhibition with and without 2C4 treatment? The process-based model has been programmed with known regimes of function in either sensitive or resistant modes before data generation. This important question we seek answers to is “is it possible to infer and identify regime difference (in terms of drug sensitivity and resistance to RTK inhibition) purely from the time-series data supplied to our modelling framework?” In this way we use the process-based model as a surrogate experimental system where - importantly - we are certain of the process-based model functioning and its response to drug action, but base our analyses only on the time series data derived from that model.

6.3.2 The datasource model of input data: the PI3K / PTEN / AKT signalling networks

First we refer to the work of Goltsov et al. (Goltsov A. & Harrison 2011) which hypothesised that the sensitivity-to-resistance transition may be a result of transition from compensatory features inherent in signalling networks. Having previously shown that PTEN plays a key role in the development of resistance to RTK inhibition (Faratian D. & Harrison 2009) they performed both experimental and theoretical studies to detail the behaviour of the signalling network in relation to acquisition of

drug resistance exploring the observed effect of PTEN expression levels on resistance. Their results revealed several compensatory mechanisms through which a particular behaviour could be effected through multiple cross-talk across multiple pathways and mutation drug targets leading to the development of resistance in phenotypes under specific drug regimes (Araujo R.P. 2007, H 2007, H 2004). Using computational and experimental methods they elucidated and established the link between properties of the PI3K/PTEN signalling cycle and cellular response to RTK stimuli mediated through the ERK/PI3K/PTEN/AKT pathways, determining the change in sensitivity of AKT activation to RTK inhibition by pertuzumab using a control kinetic parameter that encapsulates the functional properties of key signalling components that regulate enzyme activities (Goltsov A. & Harrison 2011). By varying the control parameter they were able to effect combination of perturbations to the relevant component, i.e. the PI3K/PTEN/AKT signalling cycle (see subsystem framed in figure 6-2), and adjust network dynamics to effect sensitivity-to-resistance transitions between different impact of inhibition by external inhibitors (Goltsov A. & Harrison 2011). The dynamics of the subsystem was characterised the control parameter

$$Ctrl = \frac{V_{PTEN}}{V_{PI3K} \cdot V_{AKT}} \quad (6.1)$$

where V_{PTEN} , V_{PI3K} , V_{AKT} are determined based on the initial concentrations $PTEN_0$, $PI3K_0$, AKT_0 , respectively, in terms of the following expressions:

$$V_{PTEN} = \frac{k_{cat.PTEN} PTEN_0}{K_{m.PTEN}}; \quad (6.2)$$

$$V_{PI3K} = \frac{k_{cat.PI3K} PI3K_0}{K_{m.PI3K}}; \quad (6.3)$$

$$V_{AKT} = \frac{AKT_0}{K_{d.AKT}}. \quad (6.4)$$

With the control parameter *Ctrl* they were able to appropriately switch on and steer some of the key features of the subsystem of the signal transduction system to effect desired changes into the systems by switching the model circuit into either sensitivity

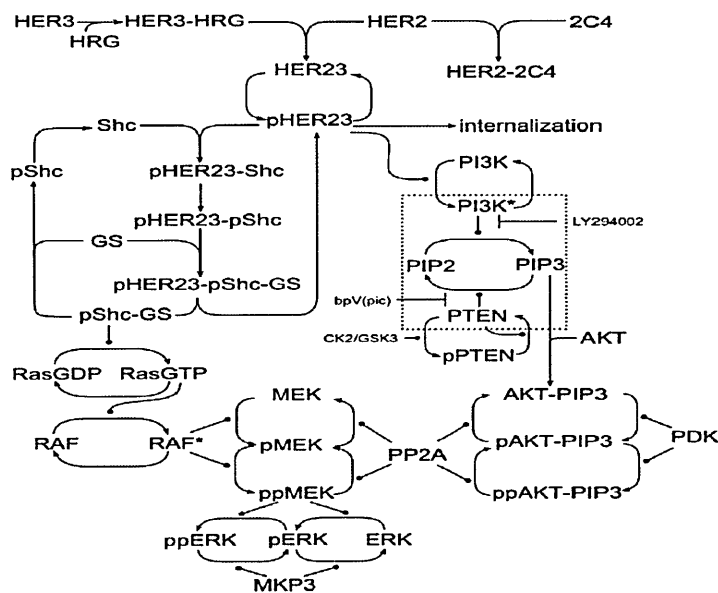


Figure 6-2: Schema of process-based model of RAF/MEK/ERK and PI3K/PTEN/AKT signaling network.

or resistant mode with or without 2C4 input. For more information see (Goltsov A. & Harrison 2011).

In view of the above simulation method for generating time series data we considered the time series data generated by Goltsov et al. (Goltsov A. & Harrison 2011) and acquired from this model output a number of samples of time series data sets all simulated from his predictive models (see figure 6-2) and used those data as data input to this case study for the assessment of our inference methods.

1

6.3.3 The acquired data samples

The time series data sets used in this case study are acquired from the process-based model developed and presented by Goltsov et al. (Goltsov A. & Harrison 2011). As

¹A subsystem of PI3K/PTEN/AKT cycle is marked by dotted frame. LY294002 and bpV(pic) are inhibitors of PI3K and PTEN, respectively. HER2, HER3: the epidermal growth factor receptors; HRG: heregulin; 2C4: pertuzumab; HER23: HER2/HER3 heterodimer; Shc: Src-homology and collagen domain protein; GS: Grb2-SOS complex; PIP2: phosphatidylinositol; PIP3: phosphatidylinositol-3,4,5-trisphosphate; RAF* and PI3K*: activated RAF and PI3K enzymes; PP2A: protein phosphatase 2A; PDK: phosphoinositide-dependent kinase 1; MKP3: MAPK phosphatase 3 (Goltsov A. & Harrison 2011).

Image credits: A. Goltsov et al. / Cellular Signalling 23 (2011) 407-416.

described in their paper, the control kinetic parameter in their model was used to serve as a resistance factor that calibrates the model's resistance responsivity measures against RTK inhibition. The data sets are supplied and grouped into four main categories, namely: S-, R-, S+, R+ representing

1. S-: systems characterised by sensitivity to RTK signals without 2C4 treatment,
2. R-: resistance to RTK signals without 2C4 treatment,
3. S+: sensitivity to RTK signals with 2C4 treatment, and
4. R+: resistance to RTK signals with 2C4 treatment, respectively.

In these models, sensitivity or resistance is determined by PTEN expression level. In order that we did not bias our data driven modelling with this key information, PTEN is not included in the time-series data set. However, the following assumptions were made during the simulation process: a) that all S- and S+ data sets assume PTEN expression level is full, hence they are regarded as being “100% *normal*”; and b) that all R- and R+ data sets assumed reduced and mutated PTEN expression level, i.e. “50% *mutated*”, hence any data generated and grouped into either the category of R- or R+ represents a cancer cell system.

6.3.4 Pertuzumab (2C4): a monoclonal antibody that targets the HER family of cell-surface receptors

The human epidermal growth factor receptor HER-2 is an important receptor for the classification of potentially aggressive breast cancer and key target in the treatment of HER-2-expressed breast cancers. Both HER-2 protein overexpression and gene amplification have been identified as key indicators of invasive breast cancer. In the treatment of breast cancer with drug agents, e.g. trastuzumab and pertuzumab, identifying the various roles that potential biomarkers associated with drug sensitivity and resistance play is key. To apply and assess the performance of our computational

and modelling method we use time series data generated from biologically plausible process-based model of signal transduction network.

As described in Chapter 4, data-consistent ODE models may be constructed from time series data that approximate or represent biological systems. We apply the TRM algorithm to the time series data supplied by Goltsov which represent the signal transduction systems represented in (Goltsov A. & Harrison 2011) in which kinetic parameters and topological information had been tested and well-defined.

We first propose that given that only time series data sets of some system measurables, akin to what might be carried out in a wet-lab experiment, we might be able to take such complex data set representative of cell lines that may or not respond to drug action and apply our inference method to identify and differentiate between those data in the [S-] and [R-] group and those in either [S+] or [R+]. If our assessment result is conclusive then the proposed method is important to any drug screening programme, i.e. based on some fairly routine biological analyses and immediate systems-scale modelling of biological data we can screen whether a drug is effective or not.

This case study primarily focuses on the assessment and capability of our inference method in differentiating between time series data representative of different system. We focus on whether or not unique signals may be identified in each of the systems which uniquely separates from any other system(s) to meaningfully indicate drug sensitivity and drug resistance.

Instant visualisation based on heatmap representation is first performed on the given time series data sets, see Figures D-1 to D-3. The higher the entry-value the deeper the colour shade that is used to depict that value. Please note that the diagonal entries are not included in the heatmap program. Such instant visualisation of data help provide instant analyses on the magnitudes of the time series data values to determine the maximum and minimum entries in a given data set. Please note that the heat maps are not calculated based on global scales so heatmap image is based on the maximum entry in each individual time series data table.

Each of the figures represents a collection of the S-, R-, S+, and R+ signal trans-

duction network of the HER2/3-PI3K-MAP signalling pathways over a certain period of capture, i.e. 8, 10, and 12 minutes, with their data in normalised form. We speculate that the information extracted about the most immediate system's response to treatment, drug's normal mechanism of action, and after-effects of 2C4 could be different depending on the time step (here 8, 10 or 12 minutes), and so the information extractable from these inference activities give insights into the robustness of our inference methods. Normalising the time series data across the different groups of data sets eases the comparison process component-wise. Finding efficient inference algorithms that are stable and reliable is difficult and it is important to demonstrate that this method is capable of making precise predictions in changing conditions.

Each time series data contains exactly 8 time points at regular intervals. Irregular time intervals within the data sets will definitely complicate the model construction problem. The measure of model sensitivity considered here is simple: that is to check whether or not the key signals expected will be shown in the interaction maps. In addition, varying the time scales might also help explore whether a particular signal would be sustained, drained, or recovered after a longer period of time or not.

Since the number of time points is less than the number of measurables within the system we also deal with a further complication in parameter estimation, akin to normal expectation in real-life scenarios. In Figures D-1 to D-3 top datasets S- and S+ (i.e top-left and top-right, respectively) represent data representative of cell lines that are potentially sensitive to RTK inhibition in the absence and presence of 2C4, respectively. The bottom datasets R- and R+ (i.e bottom-left and bottom-right, respectively) represent data representative of cell lines that have potentially developed resistance to RTK inhibition in the absence and presence of 2C4 input, respectively.

6.3.5 Modelling of *in-silico* experimental time series data

Usually the inference algorithms introduced in chapter 4 are sufficient for dealing with most of our system identification and parameter estimation challenges; only in rare cases should we require additional techniques to be used. The only exception

to purely using the transposive regression method TRM (introduced in chapter 4) is when improper parameters occur after the initial normal estimation attempt, i.e. if a single parameter is estimated with an inappropriate magnitude (e.g. 40 times more than that of every other parameter in the model).

Dealing with improper parameter estimates in models

Whenever the model restructure and parameter estimation output is unaccepted (e.g. the magnitude of parameter value is large and seems unrealistic) still we must avoid compromising on the predictive capability of the TRM method. We compromise the data-driven model reconstruction strategy which is based on the TRM algorithm if we tamper with any of the estimated values of the parameters of the model. However, if improper values of parameters are returned after the model reconstruction we may be able to pin-point specific compartments of the model that may require a necessary readjustment in its structure. Note that this readjustment would be limited to preprocessing the model structure in preparation for a reestimation of all model parameters only. Our recommendation for seeking out where such readjustments might be necessary is to first simulate the given time series data, compare the simulated data with the original data, identify the components whose data are not reproducible and then reconcile the specification in those components with the improper parameter values obtained. Such readjustment of model structure aimed at data consistency *is not* more than repreparing all the model parameters for a new estimation to eliminate doubtful results, i.e. any results that include one of more unacceptable or unconvincing parameter value(s). Steps for dealing with improper parameter estimates are clearly demonstrated in Appendix E (supplementary information).

In this case study, we investigate to see if our inference algorithm are able to generate similar networks to the processes and assumptions defined in the process-based model. To conduct this scientific research, no *a priori* information is supplied before interpreting the results. In fact the simulation of time series data was done independently by another modeller. The key objective is to investigate if our methods

can help interpret model dynamics without knowing anything about how the systems worked originally. By analysing time series data generated from the process-based model we are able to inform our understanding of responses to that intervention introduced.

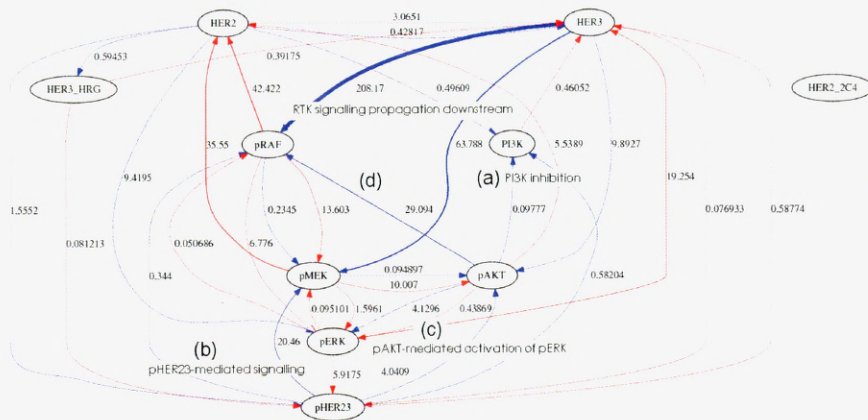


Figure 6-3: Result of S- normalised data, 8 minutes.

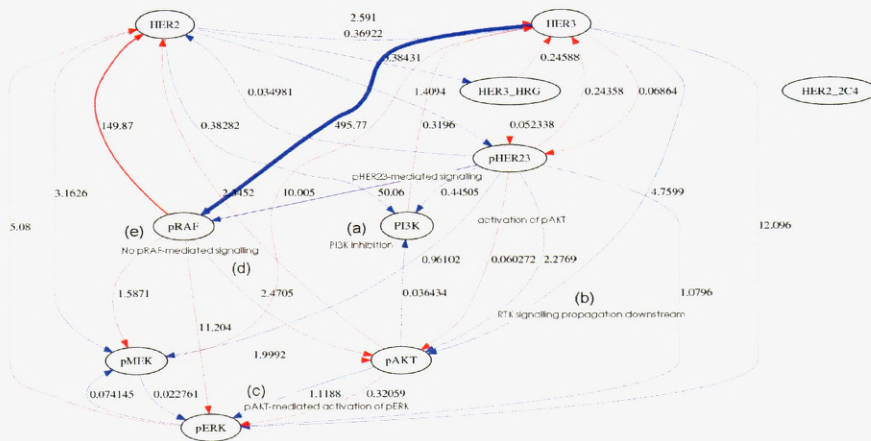


Figure 6-4: Result of S- normalised data, 10 minutes.

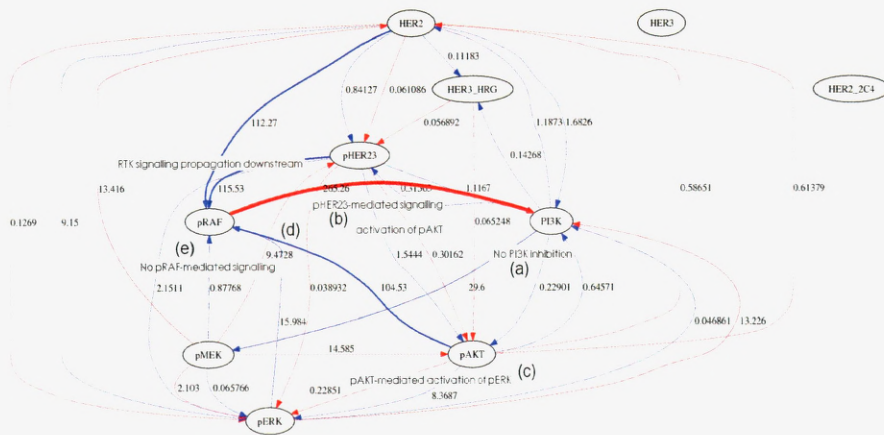


Figure 6-5: Result of S- normalised data, 12 minutes.

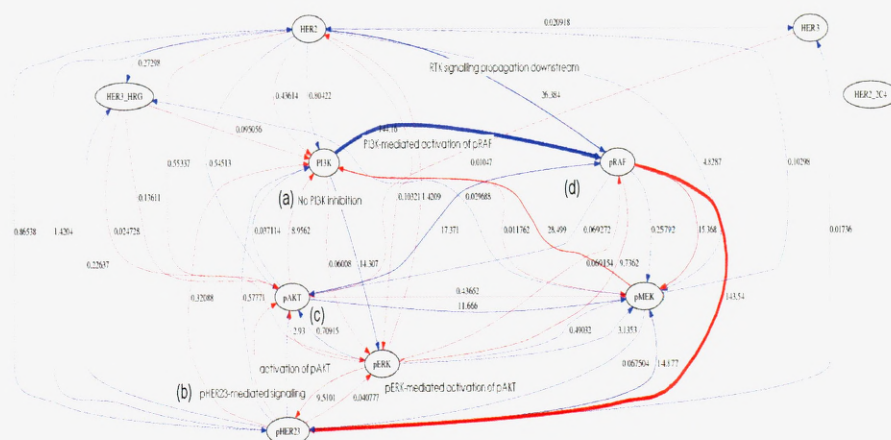


Figure 6-6: Result of R- normalised data, 8 minutes.

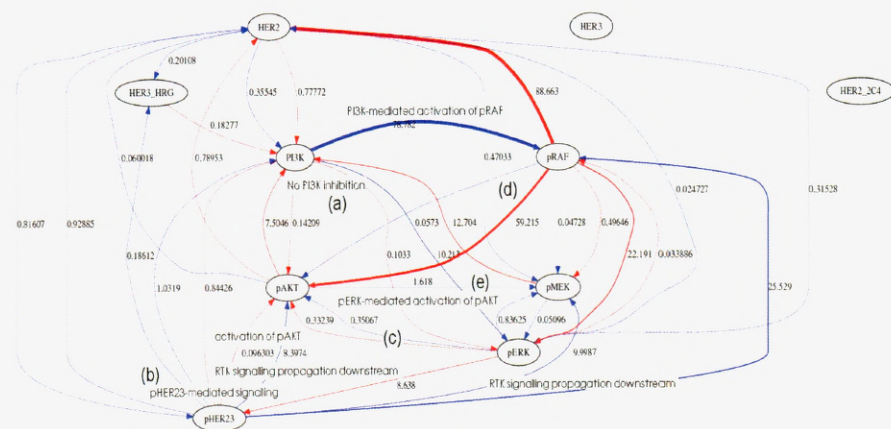


Figure 6-7: Result of R- normalised data, 10 minutes.

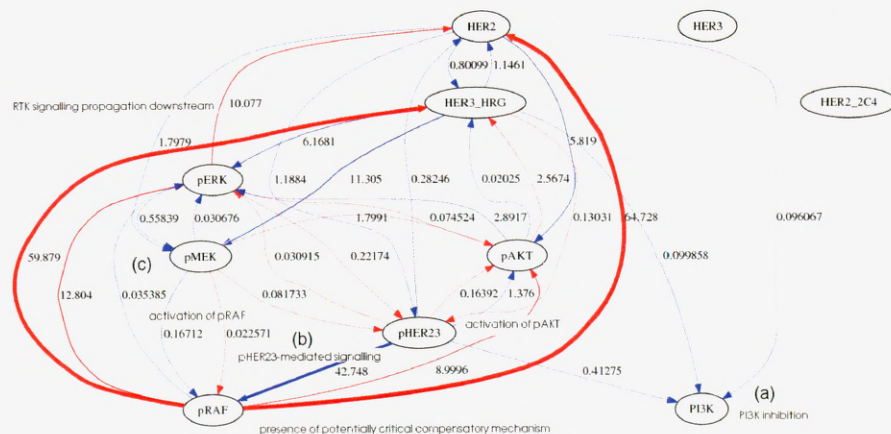


Figure 6-8: Result of R- normalised data, 12 minutes.

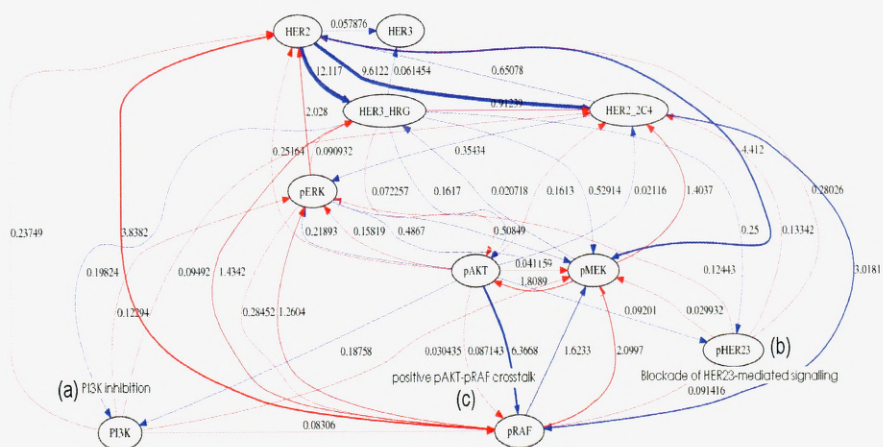


Figure 6-9: Result of S+ normalised data, 8 minutes.

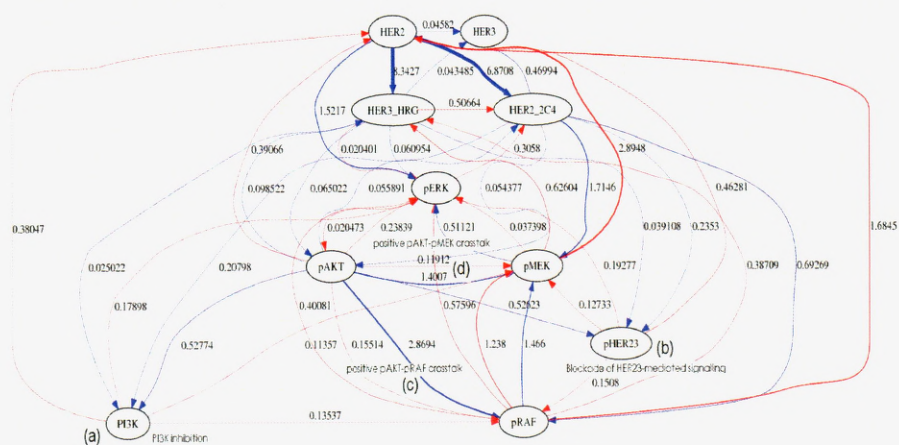


Figure 6-10: Result of S+ normalised data, 10 minutes.

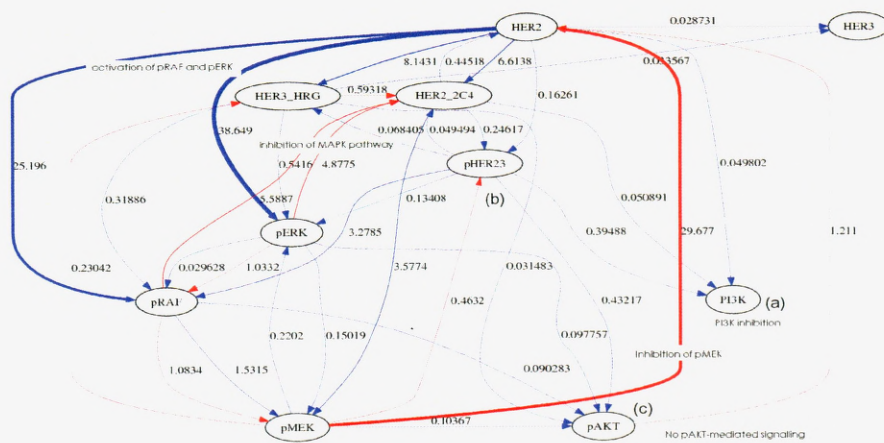


Figure 6-11: Result of S+ normalised data, 12 minutes.

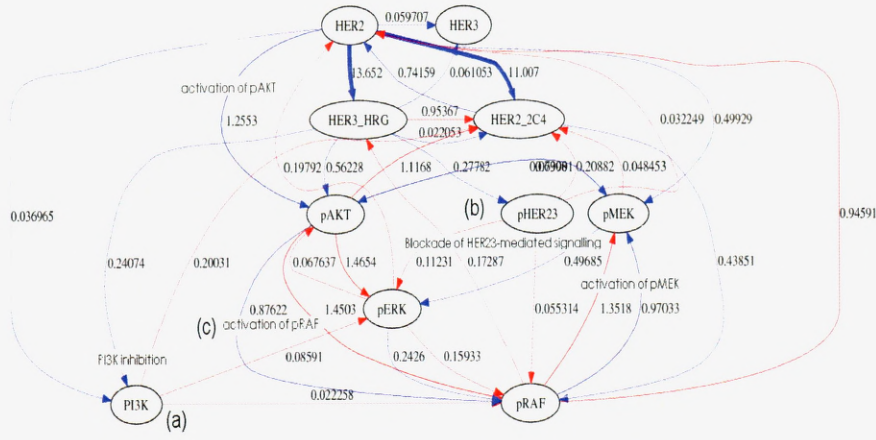


Figure 6-12: Result of R+ normalised data, 8 minutes.

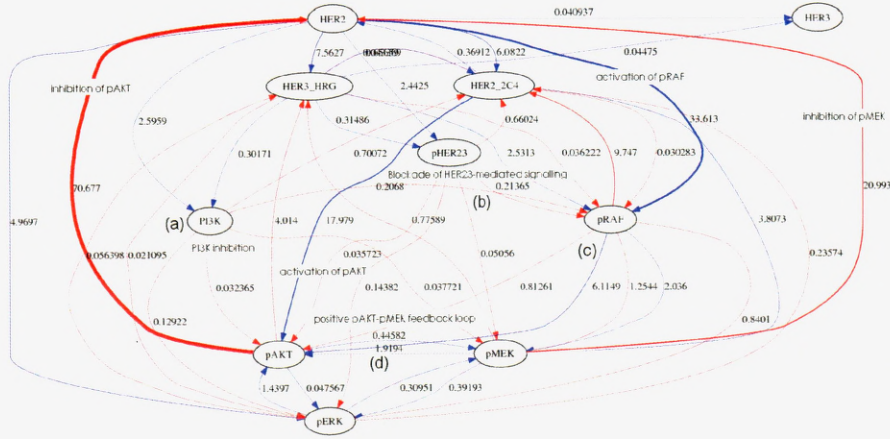


Figure 6-13: Result of R+ normalised data, 10 minutes.

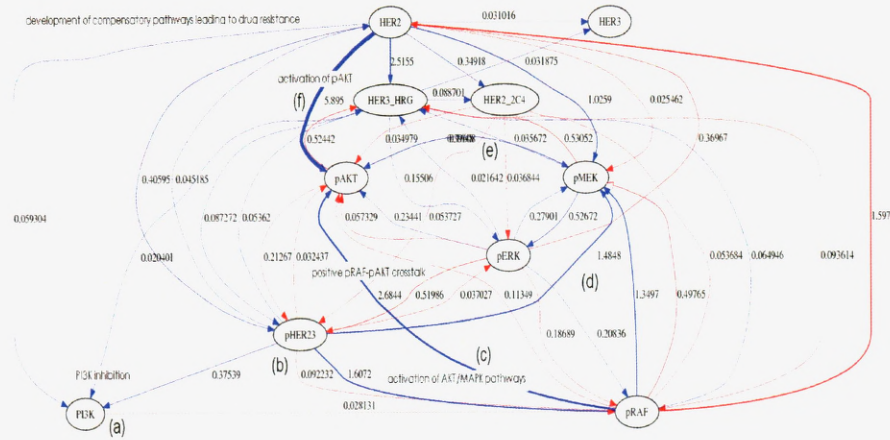


Figure 6-14: Result of R+ normalised data, 12 minutes.

6.4 Presentation of modelling and inference results

In this section we present the inference results of time series analysis. The results of time series data analyses of the normalised data are displayed as heatmaps (see figures D-5, D-6, and D-7). We search for patterns in those results, summarise some of the key features we have identified, and present the extracted features into a table (see figure 6.1 for summarised description). In the heatmaps, positive and negative signals are depicted by red (high intensity) and white (low intensity) colours, respectively.

Figures 6-3 to 6-14 represent the derived topological maps of cell lines with S- (Figures 6-3, 6-4, and 6-5), R- (Figures 6-6, 6-7, and 6-8) and R+ (Figures 6-12, S+ (Figures 6-9, 6-10, and 6-11), 6-13, and 6-14) features with 8, 10 and 12 minute time steps.

6.4.1 General consideration

Note that the results of data sets with 8 and 10 minute time steps are tagged early, while the results of those with 12 time steps are tagged as late, e.g. S- (late) or R+ (early) would represent the results of S- (8 and 10 minute time steps) or R+ (12 minute time steps), respectively. This is because for normalised data the analyses and interpretations of the topological maps are broadly independent of time step for 8 and 10 minute time steps. The data for the 12 minute time step is also similar but there are some key differences.

6.4.2 Result and interpretation

S-, R-, S+ and R+ data: activation of pAKT is indicated in all topological maps with evidence of crosstalk signals (see Figures 6-3 to 6-14).

S- and R-

Both S- and R- show the effect of PI3K inhibition by PTEN and RTK signalling mediated through HER23 dimerisation. PI3K inhibition is observed early in S- (see

Figures 6-3 and 6-4) and PI3K inhibition is indicated late in R- (see Figure 6-8), i.e. S- (8-10 minute time steps) and R- (12 minute time steps), respectively.

Note that for time step 12 minutes, PI3K is inhibited in all cases (R-, S+ and R+) except S- (see Figure 6-5). Also for 8 and 10 minute time steps, PI3K is inhibited in all cases (S-, S+ and R+) except R- (see Figures 6-6 and 6-7). Hence for time step 12 minutes, PI3K-mediated signalling is seen only in S-, and for 8 and 10 minute time steps, PI3K-mediated signalling is seen only in R-.

In both S- and R-, HER23-mediated signalling is evidenced. HER23 signalling appears to be more influential in R- than in S-. However, in R- such RTK signalling propagated downstream promotes more pAKT-mediated signalling compared to S- (see Figures 6-7 and 6-4) and positively influences PI3K-mediated signalling and crosstalk. In S- there is no evidence to suggest pRAF-mediated signalling (for 10 and 12 minute time steps). This unique feature distinguishes cell line with S- from the other cell lines, i.e. R-, S+ and R+. Another remarkable difference between S- and R- is that in R- the results reveals pRAF- or pERK-mediated crosstalk signals. No such signal is displayed in S- results.

Figure 6.1 gives a summarised description of the analyses of the derived topological maps for all time steps 8, 10, and 12 minutes.

S+ and R+

For 8, 10, and 12 minute time steps, both S+ and R+ show the effect of strong inhibition of PI3K by PTEN and 2C4, late pHER23-mediated signalling due to prolonged inhibition by 2C4 (see Figures 6-11 and 6-14), pRAF- or pERK-mediated crosstalk for 12 minute time steps (see Figures 6-11 and 6-14), pRAF-mediated signalling throughout (see Figures 6-9 to 6-14) and resultant positive pRAF-pMEK-pERK feedback loop (see Figures 6-11 and 6-14). For 8 and 10 minute time steps, in both S+ and R+ downstream propagation of HER23-mediated RTK signalling is deficient, i.e. there is no appearance of early pHER23-mediated signalling, as expected (see Figures 6-9 to 6-10 and 6-12 to 6-13). For 12 minute time steps, R+ shows pAKT-mediated crosstalk signalling (see Figure 6-14) but S+ reveals no evidence of pAKT-mediated

signalling (see Figure 6-11).

Table 6.1: Analyses of topological maps data.

cell line	PTEN inhib. seq. 8-10-12	pHER23- mediated signal. 8-10-12	pAKT- mediated crosstalk 8-10-12	pERK- or pRAF- mediated crosstalk 8-10-12	pRAF- mediated signalling 8-10-12	positive pMEK-pRAF signalling 8-10-12	nature of resultant feedback loop (for 12 min.)
S-	YYN	YYY	YNY	NNN	NNN	NNY	None
R-	NNY	YYY	YYY	YYN	YYY	NNY	pRAF \leftarrow pMEK \leftarrow pERK
S+	YYY	NNY	YYN	NNY	YYY	NNN	positive pRAF \rightarrow pMEK \rightarrow pERK
R+	YYY	NNY	YYY	NYN	YYY	NNN	positive pRAF \rightarrow pMEK \rightarrow pERK

S-, R- compared with S+, R+

HER23 inhibition is more pronounced in both S+ and R+ than in S- and R-. However, for time step 12 minutes, this inhibition of HER23 dimerisation is released (reversed) resulting in excited pRAF- or pERK-mediated crosstalk signals and positive pRAF-pMEK-pERK feedback loop in both S+ and R+ (see Figures 6-11 and 6-14). However, positive pMEK-pRAF signalling is seen in both S- and R- results for time step 12 minutes (see Figures 6-5 and 6-8).

6.5 Remarks

S-, as expected, demonstrates early inhibition of PI3K, shows no pAKT-mediated crosstalk signals for 8 minute time steps, no pRAF-mediated signalling for 10 and 12 minute time steps, and no evidence of positive feedback loop at all. R- shows little effect of PI3K inhibition for 8 minute time steps, demonstrates pAKT-mediated crosstalk signalling throughout, shows presence of pRAF- or p-ERK-mediated crosstalk for 8 and 10 minute time steps, and (reversed) positive pERK-pMEK-pRAF feedback loop. S+ may be characterised by positive pRAF-pMEK-pERK feedback loop with absence of pAKT-mediated crosstalk resulting from 2C4 input. R+ may be characterised by positive pRAF-pMEK-pERK feedback loop with presence of pAKT-mediated crosstalk resulting from 2C4 input.

These above results suggest that (sensitive) cell lines that may positively respond to 2C4 treatment are devoid of positive feedback loops and cell lines characterised by presence of pRAF- or pERK-mediated crosstalk signals are resistant and may not respond well to 2C4 treatment. With respect to response to 2C4 treatment, inhibition of HER23-dimerisation may eventually result in positive feedback loop outcome in both sensitive and resistant cell lines. However, 2C4 treatment may eventually bring about a deactivation of pAKT in the presence of PI3K inhibition in sensitive cell lines. Such a favourable outcome (inhibition of the survival of sensitive cancer cell lines) cannot be guaranteed in resistant cell lines. We detect 2C4's mechanism of action works through blockade of receptor signalling through AKT but can create a positive feedback loop in the mitogen-activated protein kinase cascade.

Clearly, network structure identification and parameter estimation of dynamical systems are necessary steps in representing system dynamics in terms of interaction networks. We demonstrated that algorithmic analysis of time series data may be very useful and could serve as both a data mining strategy and knowledge discovery method for making predictions and generating new hypotheses in complex systems modelling. On a promising note, the novel inference algorithms presented in this case study required no additional knowledge input except experimental time series data sets. As demonstrated in this case study using worked examples aimed at drug design, we here again have demonstrated the effectiveness and practicability of our inference algorithm through scientific assessment of its performances under controlled test conditions using simulated data sets generated from biologically plausible process-based models of a real system constructed by Goltsov et al (Goltsov A. & Harrison 2011).

We have demonstrated the application of our method by considering the impact of cancer drug intervention strategies on the (human) cell signalling network (for example (Goltsov A. & Harrison 2011)). This cell signalling model describes the PI3K/AKT signalling network and considers the effects of different perturbations on the network response to growth receptor inhibition. By perturbing the system with various mutations, distinct regimes of functioning were observed in the network.

Specifically regimes where the system was sensitive to intervention, where inhibition of the input signal led to inhibition of the output signal, and where the system was resistant to intervention, that is where the system was robust to input signal inhibition. Moreover, the transition between sensitivity and resistance was governed by a control parameter derived from the relative balance of the activities of three enzymes and drug interventions that target this balance may effect a shift in therapeutic resistance to therapeutic sensitivity. Our assessment of the results demonstrates the effectiveness of our inference algorithm both in performance and successful predictions of key underlying mechanisms of drug action strictly based on time series data analysis, i.e. without any additional information such as network connectivity and associations among the components of the dynamical systems.

Chapter 7

Conclusions

7.1 Confirmation of hypotheses

In the introduction of this thesis we set out the following hypotheses:

1. there exists an integrated modelling framework able to give exact representation of time series data and such techniques are sufficient to produce meaningful solutions to system identification and network inference problems;
2. the framework identified in 1 is *robust* and applicable to a wide range of limited conditions, i.e. limited data (where the number of measurables exceeds the number of time points), and beyond, i.e. surplus data;
3. the framework identified in 1 may inform experimental design and interpretation in biological systems;
4. the framework can produce an instantaneous result that indicates changes in cell signalling responses to drug action and specifically indicates sensitive and resistant signalling dynamics.

These hypotheses were confirmed in chapter 4 (hypotheses 1 and 2), chapter 5 (hypothesis 3) and chapter 6 (hypothesis 4). The hypotheses were formulated during the process of developing a data-driven modelling approach complementary to process-based modelling, i.e. our attempt to develop a robust computational and mathematical modelling framework for analysing time series data of dynamical systems.

7.1.1 Hypothesis 1

System identification problems are important problems in computer science, mathematics, and biomedicine (systems biology), mechanical engineering, financial mathematics, and so on. In dynamic modelling these are interconnected with finding optimal solutions to system identification and parameter estimation problems. Both system identification and parameter estimation problems are critical challenges often encountered in dynamic modelling and they pose difficult challenges and questions which involve finding solutions that are data consistent. Such solutions require application of theoretical, mathematical, and sometimes experimental methods to data analysis and systems of mathematical equations.

We reasoned that if we could typify an inverse problem as a reverse engineering challenge characterised by a well-defined system of differential equations and formulate and optimise this system of equations using time series data so that only most basic ODEs are expressed, the essential dynamics of such model may well represent the dynamics of the target system. If there is a match in the dynamics, the derived model is important, fundamental source of many other solutions, minable and sufficient to infer essential network of interactions.

In chapter 4, we demonstrated how the matrix-based reverse engineering and inferential procedure involves some form of mapping of the observed data to an exact reconstructed and representative network model inferred from data. The inference method always infers a jacobian model of the target system without inputting into the system *a priori* information about the architecture of the target systems.

We extended the predictive capability of the discovery strategy by increasing the number of distinct model solutions that can be provided to an inverse problem. This is achieved by developing new methods (i.e. variants of the fundamental methods which are slightly different from those fundamental core methods) to support jacobian to power-law model integration thereby ensuring that nonlinearities in complex systems may be captured and modelled using ODE models designed for formulating both simple and complex nonlinear phenomena.

The recast technique presented in chapter 4 is an integral part of a new theoretical framework developed to extend the capabilities of both our reverse engineering and current BST frameworks to support automated time series data modelling in systems biology and beyond. The modelling approach described is extremely fast, optimised, and completely data-driven. The method is generalised and applicable to any time series data of dynamical systems, i.e. with unknown underlying network of interactions. In addition, multiple data-consistent power-law (half-system) models may be inferred from such time series data without requiring *a priori* information about the architecture of the target systems.

Ultimately, we have devised and now have an important modelling framework for mining *any* single *limited* time series data by constructing a set of both multiple data-consistent jacobian models and multiple data-consistent power-law (half-system) models (solutions). The same strategy, if applied on *unlimited* (abundant) data often (i.e. > 90% of the time) guarantees that *the* actual solution to the inverse problem is found, and if not found, the suggested solution is either close to that original solution or *another* consistent solution to the problem (i.e. another *different but original system* through which the same data could have have been created).

Hypothesis 1 was confirmed in chapter 4 and has been tested and confirmed hundreds of times and may easily form a new reverse engineering theory for solving inverse problems.

7.1.2 Hypothesis 2

We define a robust inference method as an effective method that always successfully identifies or infer a system from which the (perturbed) data input could be generated even if the entire set or subsets of the data were changed. To say that the proposed inference method is robust and applicable to a wide range of limited data, we mean that the proposed method is effective, can be applied on any time series data (i.e. data insensitive), requires only minimum number (i.e. 3) of time points (i.e. 3) to infer a workable system capable of explaining the data (data consistent), and able to identify any true, unique, or optimal solution. However, this robustness is ascertained

to produce exact reconstruction of network only under some certain unrestricted conditions (exact inference): that the number of unique time points is at least equal to the number of dependent variables (system measurables); and the supplied time series data are recorded at regular time intervals. As demonstrated in chapter 4 the inference method developed is data insensitive, data consistent and capable of inferring the true (actual) system from which the time series data were generated. It is this robust feature that distinguishes out inference method from many other inference strategies. In addition, even in extreme underdetermined conditions (i.e. where the number of timepoints is between 3 and a number less than total number of measurables) the algorithms are resilient, sophisticated, still able to infer a solution that is most consistent with the limited data. Hypothesis 2 was confirmed by the results generated in chapter 4, 5, and 6. Hence we may regard this inference method as a generalised reverse engineering theory for solving inverse problems.

7.1.3 Hypothesis 3

To test our inference algorithms on real biological data, we applied it to time series (proteomic) data of DNA-damage response (DDR) signal transduction pathway. As demonstrated in chapter 5, the method when previously applied on data engaged a topological modelling of DDR and facilitated a data-rich interpolation (repopulation or refill) of the original data, without which no data-rich graph of the dynamics of DDR could be plotted under limited data availability. From those plotted graphs oscillatory patterns were deciphered as important distinguishing features inherent in the dynamics of DDR at $0.4\mu\text{M}$ Doxorubicin (Dox) treatment with and without ATM kinase inhibition. To further elucidate the mechanism behind such attribution of critically important signalling alterations, extensive analysis of DDR data at $0.4\mu\text{M}$ Dox was identified as means of differentiating between the two sets of treatments, i.e. treatments $0.1\mu\text{M}$ Dox vs. $0.4\mu\text{M}$ Dox. Such computationally determined inference and interpretation is important to channel the course of further experimental testing or better inform experimental design to verify signal to response relationship, e.g. information about the requirement to acquire and recompare DDR data between 0-8

and 8-24hr time points of treatments, the inference of the implied switch of E2F1 and chk2 influences on ATR from negative in 0-8hr to positive in 8-24hr. Hypothesis 3 was evidenced in chapter 5 informing experimental design and interpretation, suggesting how further insights may be gained from regenerated computationally developed topological map of critically important subsets of marked data.

7.1.4 Hypothesis 4

To test the inference methods on another biological system, we applied it on time series (proteomic) data simulated from biologically plausible process-based model of RTK/PI3K/AKT signalling pathways. Again as demonstrated in chapter 6, the application of the inference algorithm extends to any time series data of dynamical systems, where those systems can be identified through network inference. In that case study we were especially interested in the impact of cancer drug intervention strategies on the (human) cell signalling network using pseudo-real experimental data, i.e. data generated from well-tested process-based model of real biological systems which described the PI3K/AKT signalling network and considered the effects of different perturbations on the network response to growth receptor inhibition. By perturbing the system with various mutations, distinct regimes of functioning were observed in the network. The result of the modelling confirmed changes in cell signalling dynamics, e.g. causation of sensitivity to drug intervention, effects of inhibition of RTK signal and how it affected output signals, drug resistance mechanisms, and transition or switch mechanisms between therapeutic sensitivity and therapeutic resistance. The results of time series data analyses of the normalised data were displayed (see figures D-5, D-6, and D-7). We searched for patterns in those results, summarised some of the key features inferred from the topological maps, and presented the extracted features into a table (see figure 6.1 for summarised description). The features presented in figure 6.1 are extracted purely from the in-silico determined heatmaps presented in chapter 6 and topological maps appended in the appendix. Ultimately a table of summarised explanation about those results are presented in figure 6.1 as another justifiable proof of hypothesis 4.

7.2 Concluding remarks

Modern biology is concerned with the understanding of the structures of biological systems both at the systemic and molecular levels. In gaining this understanding the key natural phenomena involved in biological growth, evolution and processes must be understood. This has always been a challenging process depending on the level of limitation imposed on the data capture mechanism or inference method. Notwithstanding, most of the basic functions of biological components and mechanisms of their fundamental processes, which are involved at the genetic, cellular, and organic level, together with both favourable and unfavourable conditions that affect or determine their overall responses and behaviours are now being studied and understood at a scale more than ever imagined before. However, the strategic method for studying cancer biological systems requires more than just an hybridization of the best conventional reductionist approach or most effective holistic approach. Whichever approach that is being used must take cognisance of essential fundamental needs such as instant system identification requirements, fast model construction, data consistency, optimal utilisation of limited data, accurate forecasting, and new knowledge discovery (Bansal M. 2007, Gennemark P. 2009, Shovman M. 2010, Idowu M.A. 2011*b*, Idowu M.A. 2011*a*, Idowu M.A. 2012, Bown J. 2012, Idowu M.A. 2013).

In this thesis, the requirement to capture and model system dynamics before inferring topological features is first recognised (Albert R. 2001). The inference method that has been developed adopts a generic approach to dynamic modelling in such a way that *any* time series data of complex system (e.g. biological systems and beyond) can be modelled with seconds or a minute. We identified the need to provide an inference system that is able to support both systemic level modelling and understanding (Torres N.V. 2003) and dynamic modelling of subsystems based on availability of experimental time series measurements. Our dynamic modelling strategy targets the need to provide techniques that can support structural and dynamic analyses (Andrea Sackmann & Koch 2006). Incorporated into our modelling framework is a visualisation pipeline for generating *in silico* topological map or network diagram

of a constructed model in 2D using Graphviz tool. In building a model of artificial networks we first used assessment tests to develop, select, refine and optimise our inference methods based on qualitative and quantitative analyses of results generated by those methods. The condition to avoid using any *a priori* information about the original networks of interactions of the target systems was identified and satisfied throughout the simulation experiments and all case studies. Not only that, we ensured that none of the methods we have developed uses an iterative procedure to estimate parameters (Kitayama T. 2006). Our inference algorithm is able to construct both the jacobian or power-law based model that is consistent with the given time series data (Goel G. 2008) using the least number of ever parameters possible. To satisfy the requirement to optimally utilise limited data as much as possible, the strategy we have used is able to construct a workable dynamic model from subsets of large quantities of time series data or limited data, including those with a minimum of three (3) time points. The reverse engineering solution we have provided is matrix-based and uses a deterministic modelling approach. Hence we are able to provide matrix-based analytical methods and recast techniques (Shovman M. 2010, Idowu M.A. 2011b, Idowu M.A. 2011a, Idowu M.A. 2012, Idowu M.A. 2013) for transforming to and from one model type (e.g. jacobian model) to another (e.g. power-law model). Ultimately we have provided an inference method for solving inverse problems based on time series data. Our method does not require high computing power to work because it is based strong mathematical analysis which makes it highly convenient and desirable to use.

In summary, although most of the initial development and assessment tests were based on computational studies, at this maturing phase nearly all the fundamental results (after eliminating all heuristic approaches and trivial solutions) are fundamentally based on theoretical and mathematical analyses. The computational aspects of the approach have now been trivialised by the powerful mathematical analyses that have been founded on basic matrix operations. Hence the algorithm runs in seconds.

7.3 Future work and considerations

A method for reverse engineering limited and unlimited time series data can be useful for a number of reasons. The following progress and limitations are worthy of note.

1. It is widely known that the solution set of a system of ordinary differential equations can be represented by any combination of exponential functions of eigenvalues and their eigenvectors, I introduced and demonstrated its application to data discretisation using matrix-based technique (Idowu M.A. 2011a). The reverse engineering technique introduced in (Idowu M.A. 2011a) may further be developed into a sophisticated quantisation algorithm (quantiser) in such a way that large quantities of time series data may be *quantized* into a simple model that represents those data to save space.
2. I introduced a new approximation technique for calculating the logarithmic inverse of a matrix for the first time in (Idowu M.A. 2011a) using a scaling factor μ . What is the best way to determine the optimal range for μ .
3. Are there better ways to improved the layout of the *in insilico* topological maps presented in chapters 5 and 6?
4. I developed and presented a new method for constructing and decomposing square matrices (see appendix G (Idowu M.A. 2012)) which is useful for creating the non defective transformation matrices that were used in data discretisation and benchmark tests described in chapter 4. Currently the decomposition algorithm is being considered for decompartmentalisation of “big” models into smaller decompartmentalised model subunits. What is the best compartmentalisation theory available for this?
5. I demonstrated that the matrix factorisation algorithm presented in appendic (Idowu M.A. 2011a) is fundamentally well-connected to the Cholesky decomposition if applied on symmetric matrices. I also demonstrated that it is related to the LU decomposition method via a diagonal matrix multiplier. Through the

method a new direct relation between Cholesky decomposition and LU factorisation is demonstrated for the first time in (Idowu M.A. 2011a). The original purpose for creating this matrix factorisation technique is: a) to provide support to find the various associations between the predetermined partial derivatives of a jacobian model and other unknown partial derivatives within the same model and b) to promote and report parametric reverse engineering. A theoretical foundation for parametric reverse engineering has been laid in the form of matrix factorisation based on their implied minors. However, this still requires some groundbreaking research to take root.

Apart from the above-mentioned considerations, there are more immediate issues to consider. The following are pressing challenges in this thesis: considerations of ill-posed (bad data) inverse problems; need to improve current technique to accommodate data with irregular time intervals; determination of factors influencing non identifiability of a unique solution under surplus data; challenges posed due to linear dependency in time series data; recasting a jacobian model to process-based model with Michaelis-Menten formalism; design of web based user interface for automated reverse engineering; development of standalone user interface for 2D/3D visualisation of topological maps and state transition results; application of method to geneomic data, proteomic data, tissue data, patient data; application of method to non biological system modelling.

The inference algorithms we have developed may be used to increase current understanding of biological systems at system levels through application on genomic, proteomic, tissue-level and other personalised data of individual patients. Since the time-scales associated with most of these multi-scale and multi-process systems can drastically differ from one another, a time-scale insensitive inference approach is required for processing their data based on multiple temporal scales. Such is our inference method that it can process multi-scales and is applicable to a wide range of time series data dictated by highly complicated factors, conditions, and environments, e.g. many snapshots of time-based responses driven at different scales and influenced by spatial factors. The solutions provided to these inference challenges

may be very useful and complementary to process-based modelling approaches, e.g. they may provide relevant and useful meta view on how to constrain process-based modelling of multi scale systems. It may be possible to incorporate the inferential power and capability of the novel data-driven modelling strategy into a process-based modelling framework of another's gearing the integrated framework towards instant inference of data and promoting system level modelling and understanding. Imagine a combined system-wide mapping content based on topological maps of network of interactions inferred from genomic, proteomic, and tissue-level time series data all providing the much required information about target cellular systems. The impact that such innovation could make would be huge in furthering scientific wisdom and contemporary understanding of biological sciences.

Appendix A

Figures

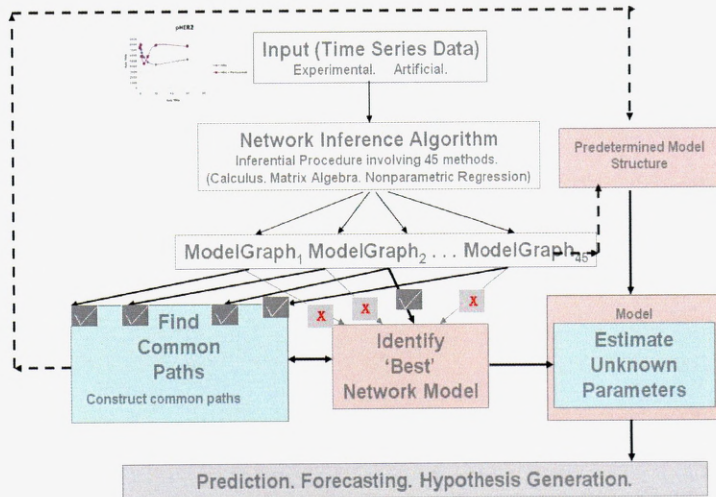


Figure A-1: The initial proposed system identification (inference) framework.

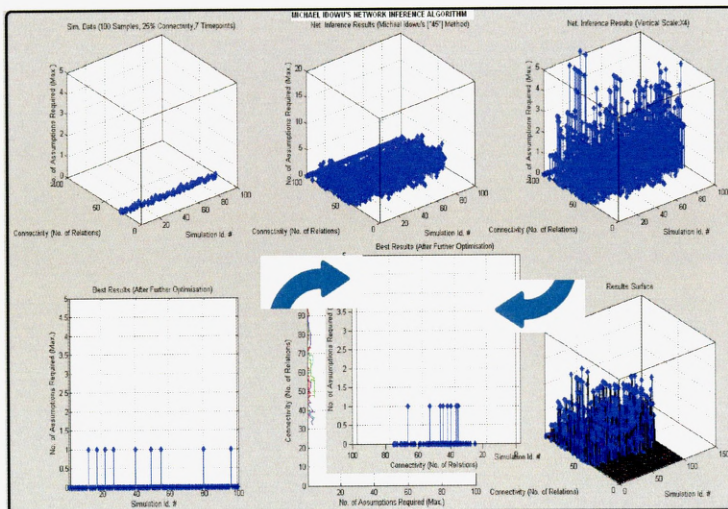


Figure A-2: Evaluation of optimum results (100 networks).

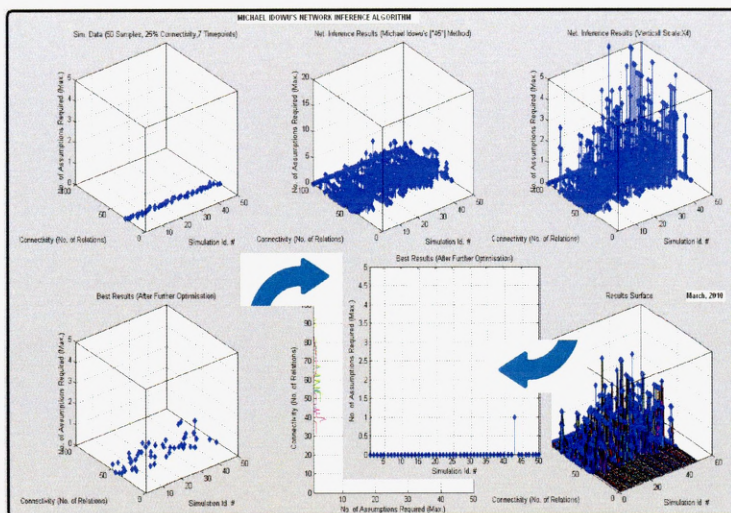


Figure A-3: Evaluation of optimum results (50 networks).

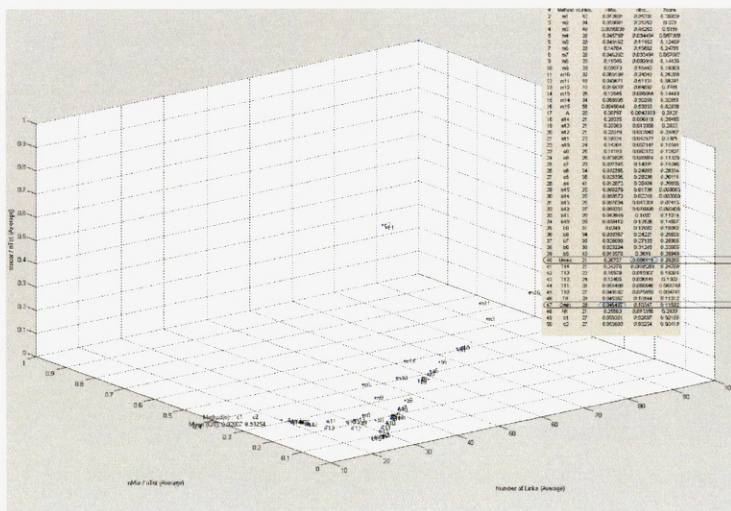


Figure A-4: Assessment and comparison of optimal system identification methods

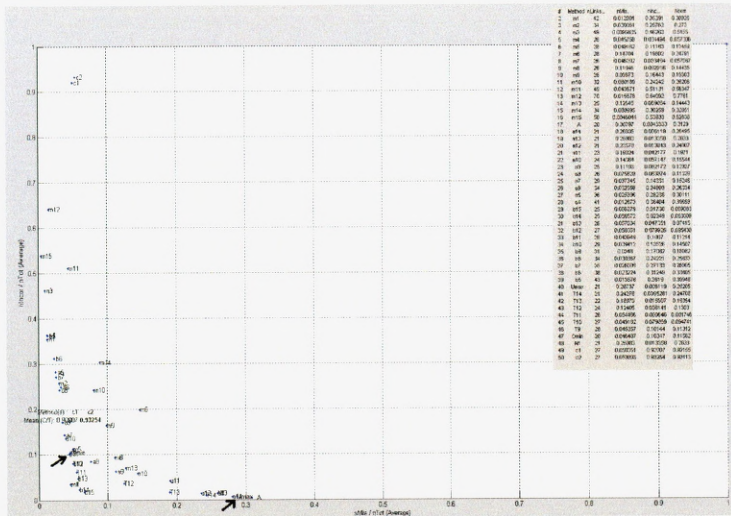


Figure A-5: Assessment and comparison of optimal system identification methods

#	Method	nLinks...	nMis...	nInc...	Norm
2	m1	42	0.012801	0.35391	0.38828
3	m2	34	0.030061	0.25762	0.273
4	m3	49	0.0096635	0.46263	0.5155
5	m4	26	0.045758	0.034494	0.057306
6	m5	28	0.049162	0.11163	0.12459
7	m6	28	0.14704	0.19802	0.24791
8	m7	26	0.046202	0.033494	0.057067
9	m8	26	0.11046	0.092916	0.14435
10	m9	28	0.09873	0.16443	0.19303
11	m10	32	0.080189	0.24242	0.26206
12	m11	49	0.043671	0.51131	0.56347
13	m12	70	0.015578	0.64092	0.7761
14	m13	25	0.12645	0.069064	0.14443
15	m14	34	0.088995	0.30259	0.32651
16	m15	58	0.0046044	0.53933	0.62836
17	A	20	0.30797	0.0043333	0.3129
18	a14	21	0.28035	0.006119	0.28495
19	a13	21	0.25883	0.013358	0.2633

38	b6	38	0.023224	0.31249	0.33605
39	b5	43	0.013576	0.3619	0.39948
40	Umax	21	0.28737	0.006119	0.29205
41	T14	21	0.24276	0.0095281	0.24708
42	T13	22	0.18979	0.015907	0.19354
43	T12	24	0.12405	0.036141	0.1303
44	T11	26	0.054486	0.060646	0.081748
45	T10	27	0.049182	0.079859	0.094741
46	T9	28	0.045357	0.10144	0.11312
47	Omin	28	0.046487	0.10347	0.11532
48	N1	21	0.25883	0.013358	0.2633
49	c1	27	0.050351	0.92007	0.92155
50	c2	27	0.053803	0.93254	0.93413

Figure A-6: Optimum result selection: “best overestimates” and “best underestimates” .

Appendix B

Figures

Experimental time series data (Fold induction - 0.1μM DOX)								
	pATM	pATR	pP ₅₃	pChk ₂	pChk ₁	pBRCA	pE ₂ F1	pH ₂ AX
pATM →	-0.0063639	0.021938	0	-0.023432	0	-0.0074367	0.010039	0.003382
pATR →	0.0055985	-0.025935	0	0.031987	0	0.0078203	-0.011401	-0.010265
pP ₅₃ →	1.461	-1.4586	-0.12558	-0.010072	-0.017642	0.76363	-0.31173	-0.5777
pChk ₂ →	-0.0037111	0.029526	0	-0.040635	0	-0.0085718	0.014513	0.011679
pChk ₁ →	0.98043	-1.035	0.01205	0.11735	-0.15298	0.53081	-0.22838	-0.41967
pBRCA →	0.017408	-0.049562	0	0.047529	0	0.017431	-0.017971	-0.019272
pE ₂ F1 →	-0.010681	0.094843	0	-0.13173	0	-0.025128	0.042832	0.03672
pH ₂ AX →	0.012177	0.0055705	0	-0.029756	0	0	0.010989	0.0024956

Experimental time series data (Fold induction - 0.4μM DOX)								
	pATM	pATR	pP ₅₃	pChk ₂	pChk ₁	pBRCA	pE ₂ F1	pH ₂ AX
pATM →	-0.0063639	0	0.0025298	0	0.0011134	0	0	-0.001931
pATR →	0.011699	-0.025935	0.0028373	0	-0.0032131	-0.026218	0.0040806	-0.001325
pP ₅₃ →	0.63277	0	-0.12558	0	0.20594	0.17037	-0.47364	0.29619
pChk ₂ →	0.0085523	0	0.006882	-0.040635	0.027043	0.019041	-0.023152	0.010122
pChk ₁ →	-0.013528	0	-0.027299	0	-0.15298	-0.19142	0.28038	-0.076008
pBRCA →	0.0053829	0	0	0	0.0096654	0.017431	-0.016992	0.0039567
pE ₂ F1 →	-0.0084069	0	-0.0084341	0	-0.0226	0.0098278	0.042832	-0.011893
pH ₂ AX →	0.022069	0	-0.0066806	0	-0.006077	-0.030188	0.01033	0.0024956

Experimental time series data (Fold induction - 0.1μM DOX +U)								
	pATM	pATR	pP ₅₃	pChk ₂	pChk ₁	pBRCA	pE ₂ F1	pH ₂ AX
pATM →	-0.0063639	0	-0.0040328	0	0.016946	-0.018553	0.015461	-0.0029364
pATR →	0.0013605	-0.025935	0.0011077	0.002266	-0.0014289	0	0.023673	0
pP ₅₃ →	-0.10018	0	-0.12558	0	0.082459	-0.018434	0.20878	-0.036912
pChk ₂ →	0.0024083	-0.0023287	0.030381	-0.040635	-0.011135	0.016491	-0.026549	0.023573
pChk ₁ →	0.084618	0	0.14516	0	-0.15298	0.089001	-0.21974	0.02202
pBRCA →	0.010613	0	0.022754	0	-0.024872	0.017431	-0.031503	0
pE ₂ F1 →	-0.015285	0	-0.000277	0	0.021682	-0.0021436	0.042832	-0.013363
pH ₂ AX →	0.049756	0	0.18422	0	-0.31721	0.001018	-0.27076	0.0024956

Experimental time series data (Fold induction - 0.4μM DOX +U)								
	pATM	pATR	pP ₅₃	pChk ₂	pChk ₁	pBRCA	pE ₂ F1	pH ₂ AX
pATM →	-0.0063639	-0.0010972	0.0025268	-0.0018784	0.0025482	0	-0.0018281	0
pATR →	-0.057925	-0.025935	0.045179	0.041613	0.01828	0	-0.023885	0
pP ₅₃ →	0.36816	0.44362	-0.12558	-0.84685	-0.16997	0	0.082574	0
pChk ₂ →	0.031285	0.009331	-0.0078869	-0.040635	-0.0073899	0	0.007823	0
pChk ₁ →	1.0408	0.19931	-0.19341	-1.1183	-0.15298	0	0.12108	0
pBRCA →	-0.023283	0.0026945	0.0022074	-0.014891	-0.0027223	0.017431	0.0071906	0
pE ₂ F1 →	0.059213	0.034711	-0.026419	-0.11919	-0.030369	0	0.042832	0
pH ₂ AX →	-0.003799	0.0054085	-0.0018617	-0.016545	-0.0048936	0	0.0084594	0.0024956

Figure B-1: The equivalent derived half-system representations of the four systems.

Appendix C

Tables

Table C.1: S-, normalised data, 8 minutes. Inferred matrix of signalling network

	HER2_2C4	HER3_HRG	HER_2	HER_3	pHER_23	PI3K	pAKT	pRAF	pMEK	pERK
HER2_2C4	0.11684	0	0	0	0	0	0	0	0	0
HER3_HRG	0	0.11684	0.59453	-0.39175	-0.081213	0	0.00097477	0.00031813	0.0012719	-0.00054211
HER_2	0	0	-0.062939	-3.0651	0.01887	0	0.010622	-0.00018926	0.00011554	0.005046
HER_3	0	0	0.42817	-5.9601	-0.076933	0	0.0033227	0	0.0016198	0.00075748
pHER_23	0	0	1.5552	-0.58774	-0.35961	0	-0.041841	-0.0017727	0.0020644	0.0098105
PI3K	0	0	0.49609	-0.46052	0.58204	0.11684	0.09777	0.0013574	0.0048934	0.010772
pAKT	0	0	-5.5389	9.8927	4.0409	0	-2.349	0.012564	0.094897	-0.43869
pRAF	0	0	-42.422	208.17	0.344	0	29.094	-0.98067	-13.603	-6.776
pMEK	0	0	-35.55	63.788	20.46	0	-10.007	0.2345	-0.25296	-1.5961
pERK	0	0	9.4195	-19.254	-5.9175	0	4.1296	-0.050686	-0.095101	-0.37466

S-, normalised data, 8 minutes.

Table C.2: S+, normalised data, 8 minutes. Inferred matrix of signalling network

	HER2_2C4	HER3_HRG	HER_2	HER_3	pHER_23	PI3K	pAKT	pRAF	pMEK	pERK
HER2_2C4	-1.2295	0.017084	9.6122	0	0	0	0.02116	-0.006245	-0.0032535	-0.0097418
HER3_HRG	-0.91239	0.043299	12.117	0	0	0	-0.072257	0.0057339	0.020718	0.0056158
HER_2	0.65078	0.0010841	-11.218	0	0	0	0.0016816	-0.00047422	-0.00026663	-0.00073892
HER_3	-0.0073602	0.061454	0.057876	0.11684	0	0	-0.0015009	0	0.00026629	-0.00030834
pHER_23	-0.13342	0.25	-0.28026	0	0.11684	0	0.09201	-0.091416	-0.029932	-0.12443
PI3K	-0.09492	0.19824	-0.23749	0	0	0.11684	0.18758	-0.08306	-0.030435	-0.12294
pAKT	-0.1613	0.1617	-0.25164	0	0	0	0.032715	-0.087143	-0.041159	-0.15819
pRAF	3.0181	-1.4342	-3.8382	0	0	0	6.3668	-1.572	-2.0997	-1.2604
pMEK	-1.4037	0.52914	4.412	0	0	0	-1.8089	1.6233	0.28323	0.50849
pERK	0.35434	-0.090932	-2.028	0	0	0	0.21893	-0.28452	0.4867	-0.45816

S+, normalised data, 8 minutes.

Table C.3: R-, normalised data, 8 minutes. Inferred matrix of signalling network

	HER2_2C4	HER3_HRG	HER_2	HER_3	pHER_23	PI3K	pAKT	pRAF	pMEK	pERK
HER2_2C4	0.11684	0	0	0	0	0	0	0	0	0
HER3_HRG	0	0.11684	0.27298	0	0.22637	-0.095056	-0.13611	0.00063424	0.029688	-0.024728
HER_2	0	0	-1.6894	0	1.4204	-0.43614	-0.55337	0.0022856	0.10298	-0.10321
HER_3	0	0	0.020918	0.11684	0.01736	-0.0074169	-0.01047	0	0.00228	-0.0019061
pHER_23	0	0	0.86538	0	0.40916	-0.32088	-0.33091	-0.00162	0.067504	-0.040777
PI3K	0	0	0.80422	0	0.57771	-0.74491	0.037114	-0.0046761	-0.011762	-0.06008
pAKT	0	0	0.54513	0	7.0681	-8.9562	1.4628	0.069272	-0.43652	0.70915
pRAF	0	0	26.384	0	-143.54	144.16	17.371	-1.423	-15.368	-9.7362
pMEK	0	0	4.8287	0	14.877	-28.499	11.666	0.25792	-2.4907	3.1353
pERK	0	0	-1.4209	0	-9.5101	14.307	-2.93	-0.069154	0.49032	-1.9655

R-, normalised data, 8 minutes.

Table C.4: R+, normalised data, 8 minutes. Inferred matrix of signalling network

	HER2_2C4	HER3_HRG	HER_2	HER_3	pHER_23	PI3K	pAKT	pRAF	pMEK	pERK
HER2_2C4	-1.3254	0.022053	11.007	0	0	0	0.0015913	0	0.00073126	-0.0053267
HER3_HRG	-0.95367	0.0185	13.652	0	0	0	0.0029608	0.00010331	0.0020439	-0.0093586
HER_2	0.74159	0.0015144	-12.708	0	0	0	0.00011544	0	0	-0.00039812
HER_3	-0.0067761	0.061053	0.059707	0.11684	0	0	0	-0.00020418	-0.00012351	-0.00080398
pHER_23	-0.20882	0.27782	-0.032249	0	0.11684	0	-0.0045136	-0.055314	-0.015317	-0.11231
PI3K	-0.20031	0.24074	0.036965	0	0	0.11684	0.012012	-0.022258	0.0032215	-0.08591
pAKT	-1.1168	0.56228	1.2553	0	0	0	-0.22748	-1.4503	0.7908	-1.4654
pRAF	0.43851	-0.17287	-0.94591	0	0	0	0.87622	0.55676	-1.3518	0.2426
pMEK	-0.048453	-0.0016989	0.49929	0	0	0	0.063901	0.97033	-0.10644	-0.017036
pERK	0.008296	0.0089932	-0.19792	0	0	0	-0.067637	-0.15933	0.49685	-0.26601

R+, normalised data, 8 minutes.

Table C.5: S-, normalised data, 10 minutes. Inferred matrix of signalling network

	HER2_2C4	HER3_HRG	HER_2	HER_3	pHER_23	PI3K	pAKT	pRAF	pMEK	pERK
HER2_2C4	0.24967	0	0	0	0	0	0	0	0	0
HER3_HRG	0	0.24967	0.38431	-0.24588	-0.052338	0	0.0018854	0.00011924	0.0018809	0.00036178
HER_2	0	0	-0.11546	-2.591	0.034981	0	0.0069652	0	-0.0011059	0.0042589
HER_3	0	0	0.36922	-5.0968	-0.06864	0	0.0042882	0	0.0013675	0.0011904
pHER_23	0	0	1.4094	-0.24358	-0.29857	0	-0.060272	-0.0012618	0.00058773	0.003031
PI3K	0	0	0.38282	-0.3196	0.44505	0.24967	0.036434	0.00091732	0.0050355	0.0013844
pAKT	0	0	-2.3452	4.7599	2.2769	0	-1.3967	-0.0061373	0.0051843	-0.32059
pRAF	0	0	-149.87	495.77	50.06	0	-2.4705	-1.0378	-1.5871	-11.204
pMEK	0	0	3.1626	-10.005	0.96102	0	-1.9992	0.0015499	0.21988	0.074145
pERK	0	0	-5.08	12.096	1.0796	0	1.1188	0.017102	0.022761	-1.0051

S-, normalised data, 10 minutes.

Table C.6: S+, normalised data, 10 minutes. Inferred matrix of signalling network

	HER2_2C4	HER3_HRG	HER_2	HER_3	pHER_23	PI3K	pAKT	pRAF	pMEK	pERK
HER2_2C4	-1.0372	0.006509	6.8708	0	0	0	0.055891	-0.011689	-0.011222	-0.017232
HER3_HRG	-0.50664	-0.026718	8.3427	0	0	0	0.098522	0.0089149	-0.011411	-0.020401
HER_2	0.46994	0.00031782	-8.2495	0	0	0	0.0039563	-0.00077401	-0.00078002	-0.0011968
HER_3	-0.0035489	0.043485	0.04582	0.24967	0	0	0.0039627	-0.00065255	-0.0010172	-0.001334
pHER_23	0.2353	0.039108	-0.46281	0	0.24967	0	0.52623	-0.1508	-0.12733	-0.19277
PI3K	0.20798	0.025022	-0.38047	0	0	0.24967	0.52774	-0.13537	-0.11357	-0.17898
pAKT	0.054377	0.065022	-0.39066	0	0	0	0.36974	-0.15514	-0.11912	-0.23839
pRAF	0.69269	-0.38709	-1.6845	0	0	0	2.8694	-1.0822	-1.238	-0.57596
pMEK	1.7146	-0.62604	-2.8948	0	0	0	1.4007	1.466	-0.42135	-0.037398
pERK	-0.3058	0.060954	1.5217	0	0	0	-0.020473	-0.40081	0.51121	-0.46889

S+, normalised data, 10 minutes.

Table C.7: R-, normalised data, 10 minutes. Inferred matrix of signalling network

	HER2_2C4	HER3_HRG	HER_2	HER_3	pHER_23	PI3K	pAKT	pRAF	pMEK	pERK
HER2_2C4	0.24967	0	0	0	0	0	0	0	0	0
HER3_HRG	0	0.24967	0.20108	0	0.18612	-0.18277	-0.012746	0.0015931	0.005048	0.00092903
HER_2	0	0	-0.95948	0	0.92885	-0.77772	0.060018	0.0066261	0.006915	0.024727
HER_3	0	0	0.015798	0.24967	0.014604	-0.014455	-0.0010045	0.00012567	0.00039571	0
pHER_23	0	0	0.81607	0	0.7189	-0.84426	-0.096303	0.0040128	0.0049581	0.010806
PI3K	0	0	0.35545	0	1.0319	-0.86811	-0.14209	-0.0037016	-0.0082171	-0.1033
pAKT	0	0	-0.78953	0	8.3974	-7.5046	-1.0487	0.0573	0.004631	0.35067
pRAF	0	0	-88.663	0	25.529	78.782	-59.215	-1.7387	-0.49646	-22.191
pMEK	0	0	0.47033	0	9.9987	-12.704	1.618	0.04728	0.12069	0.83625
pERK	0	0	0.31528	0	-8.638	10.213	-0.33239	-0.033886	0.05096	-1.3665

R-, normalised data, 10 minutes.

Table C.8: R+, normalised data, 10 minutes. Inferred matrix of signalling network

	HER2_2C4	HER3_HRG	HER_2	HER_3	pHER_23	PI3K	pAKT	pRAF	pMEK	pERK
HER2_2C4	-1.0895	0.043369	6.0822	0	0	0	-0.0070478	-0.030283	-0.0037085	-0.015069
HER3_HRG	-0.65659	0.043343	7.5627	0	0	0	-0.0067231	-0.036222	-0.0041668	-0.021095
HER_2	0.36912	0.0032041	-6.7449	0	0	0	-0.00058675	-0.0024229	-0.00037094	-0.0011492
HER_3	-0.005519	0.04475	0.040937	0.24967	0	0	0.00010333	-0.00028207	-0.00020313	-0.00072032
pHER_23	-0.66024	0.31486	2.4425	0	0.24967	0	-0.035723	-0.21365	-0.05056	-0.14382
PI3K	-0.70072	0.30171	2.5959	0	0	0.24967	-0.032365	-0.2068	-0.037721	-0.12922
pAKT	17.979	-4.014	-70.677	0	0	0	1.1942	6.1149	1.9194	1.4397
pRAF	-9.747	2.5313	33.613	0	0	0	-0.81261	-3.5299	-1.2544	-0.8401
pMEK	3.8073	-0.77589	-20.993	0	0	0	0.44582	2.036	0.0031599	0.30951
pERK	-0.23574	-0.056398	4.9697	0	0	0	0.047567	-0.018877	0.39193	-0.25348

R+, normalised data, 10 minutes.

Table C.9: S-, normalised data, 12 minutes. Inferred matrix of signalling network

	HER2_2C4	HER3_HRG	HER_2	HER_3	pHER_23	PI3K	pAKT	pRAF	pMEK	pERK
HER2_2C4	0.31441	0	0	0	0	0	0	0	0	0
HER3_HRG	0	0.31441	0.11183	0	-0.056892	0.14268	-0.065248	0.00019541	0.00034356	-0.01411
HER_2	0	0	-1.6096	0	-0.061086	1.1873	-0.58651	0.0015693	-0.005936	-0.1269
HER_3	0	0	0.00895	0.31441	-0.0046785	0.011524	-0.0052733	0	0	-0.0011407
pHER_23	0	0	0.84127	0	-0.19078	0.31563	-0.30162	0.0014015	0.0028479	-0.038932
PI3K	0	0	1.6826	0	1.1167	-2.2399	0.64571	-0.0039602	0.0050867	0.046861
pAKT	0	0	-0.61379	0	1.5444	0.22901	-1.3792	-0.006219	-0.011409	-0.22851
pRAF	0	0	112.27	0	115.53	-265.26	104.53	-0.713	0.87768	15.984
pMEK	0	0	-13.416	0	-9.4728	29.6	-14.585	0.0058852	-0.067892	-2.103
pERK	0	0	9.15	0	2.1511	-13.226	8.3687	-0.014975	0.065766	0.33066

S-, normalised data, 12 minutes.

Table C.10: S+, normalised data, 12 minutes. Inferred matrix of signalling network

	HER2_2C4	HER3_HRG	HER_2	HER_3	pHER_23	PI3K	pAKT	pRAF	pMEK	pERK
HER2_2C4	-1.0402	0.0066441	6.6138	0	0.049494	0	0	0.0012252	0.0020391	0.0011176
HER3_HRG	-0.59318	-0.0030487	8.1431	0	0.068405	0	0	-0.00038255	0.0039523	-0.00065358
HER_2	0.44518	0.00037751	-7.8716	0	0.0034564	0	0	0	0.0001381	0
HER_3	-0.0029554	0.033567	0.028731	0.31441	0.0025998	0	0	0.00017058	0	0
pHER_23	0.24617	0.0022193	0.16261	0	-0.089915	0	0	-0.00013071	-0.0014056	-0.0012304
PI3K	0.050891	-0.0026779	0.049802	0	0.39488	0.31441	0	0.0066192	0.0096068	0.0063906
pAKT	-0.010031	0.031483	-1.211	0	0.43217	0	0.31441	0.090283	0.10367	0.097757
pRAF	-5.5887	0.31886	25.196	0	3.2785	0	0	-1.6683	-1.0834	0.029628
pMEK	3.5774	-0.23042	-29.677	0	-0.4632	0	0	1.5315	0.12728	0.15019
pERK	-4.8775	0.5416	38.649	0	0.13408	0	0	-1.0332	0.2202	-0.69077

S+, normalised data, 12 minutes.

Table C.11: R-, normalised data, 12 minutes. Inferred matrix of signalling network

	HER2_2C4	HER3_HRG	HER_2	HER_3	pHER_23	PI3K	pAKT	pRAF	pMEK	pERK
HER2_2C4	0.31441	0	0	0	0	0	0	0	0	0
HER3_HRG	0	-1.3222	0.80099	0	-0.13031	0	0.02025	0.00042818	0.0031043	0.0093713
HER_2	0	1.1461	-1.451	0	0.28246	0	-0.0032994	-0.00021265	-0.0048591	0.0011307
HER_3	0	0.012454	0.0088027	0.31441	-0.00097074	0	0.00029568	0	0	0.00013035
pHER_23	0	-0.014004	1.1884	0	-0.10566	0	-0.16392	-0.0022988	0.0013032	-0.030915
PI3K	0	0.099858	0.096067	0	0.41275	0.31441	0.0048365	0.00020501	-0.0028906	-0.0069281
pAKT	0	-2.5674	5.819	0	1.376	0	-2.5548	-0.0018314	-0.01517	-0.074524
pRAF	0	-59.879	-64.728	0	42.748	0	-8.9996	-1.1206	0.16712	-12.804
pMEK	0	11.305	1.7979	0	-0.081733	0	-1.7991	-0.022571	-0.049175	0.55839
pERK	0	6.1681	-10.077	0	-0.22174	0	2.8917	0.035385	0.030676	-0.63921

R-, normalised data, 12 minutes.

Table C.12: R+, normalised data, 12 minutes. Inferred matrix of signalling network

	HER2_2C4	HER3_HRG	HER_2	HER_3	pHER_23	PI3K	pAKT	pRAF	pMEK	pERK
HER2_2C4	0.31441	0.088701	0.34918	0	-0.053727	0	-0.034979	-0.093614	0.00098489	-0.036844
HER3_HRG	0	-0.062911	2.5155	0	0.05362	0	0.016633	0.064946	0.035672	0.021642
HER_2	0	0.016103	-2.6376	0	0.045185	0	-0.00066436	-0.018663	-0.025462	-0.0035541
HER_3	0	0.031875	0.031016	0.31441	0.0047198	0	0.00077625	0.0019234	-0.00030623	0.00068279
pHER_23	0	0.087272	0.40595	0	-0.1437	0	-0.032437	-0.092232	-0.0093118	-0.037027
PI3K	0	0.020401	0.059304	0	0.37539	0.31441	-0.013192	-0.028131	0.017523	-0.0032556
pAKT	0	-0.52442	5.895	0	-0.21267	0	0.25191	2.6844	0.1648	0.23441
pRAF	0	-0.053684	-1.5973	0	1.6072	0	-0.11349	-0.86843	-0.49765	0.20836
pMEK	0	-0.53052	1.0259	0	1.4848	0	0.39608	1.3497	-0.63648	0.27901
pERK	0	0.15506	-0.36967	0	-0.51986	0	-0.057329	-0.18689	0.52672	-0.35778

R+, normalised data, 12 minutes.

Appendix D

Figures

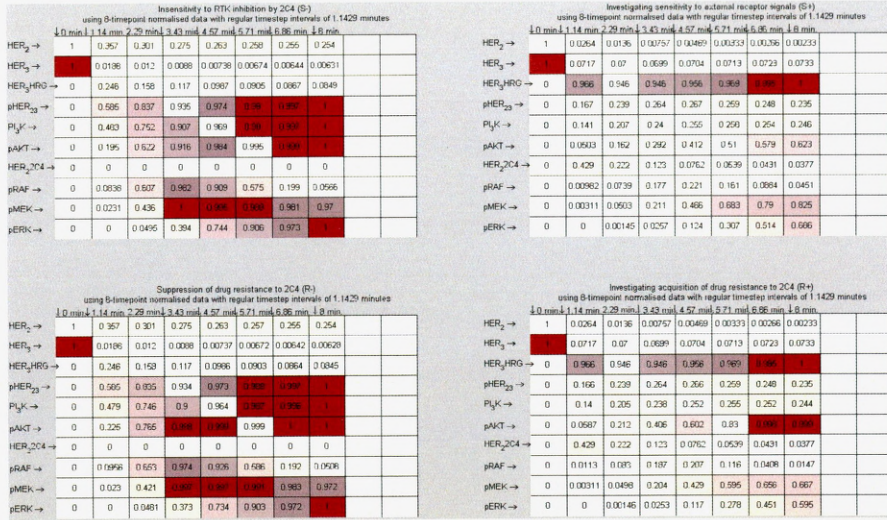


Figure D-1: Heat map representations of 4 different normalised data sets generated from a biologically plausible process-based model of HER2/3-PI3K-MAPK signalling pathways with 8-timepoint readings recorded over a period of 8 minutes only.

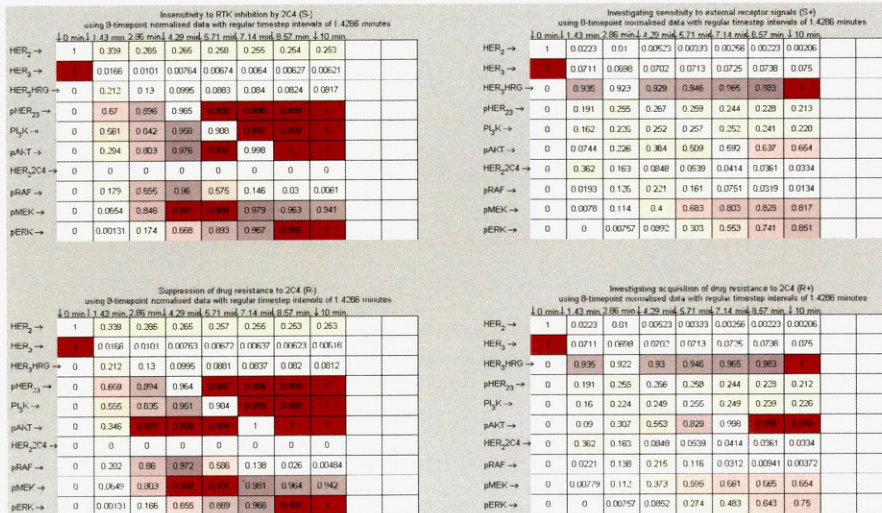


Figure D-2: Heat map representations of 4 different normalised data sets generated from a biologically plausible process-based model of HER2/3-PI3K-MAPK signalling pathways with 8-timepoint readings recorded over a period of 10 minutes only.

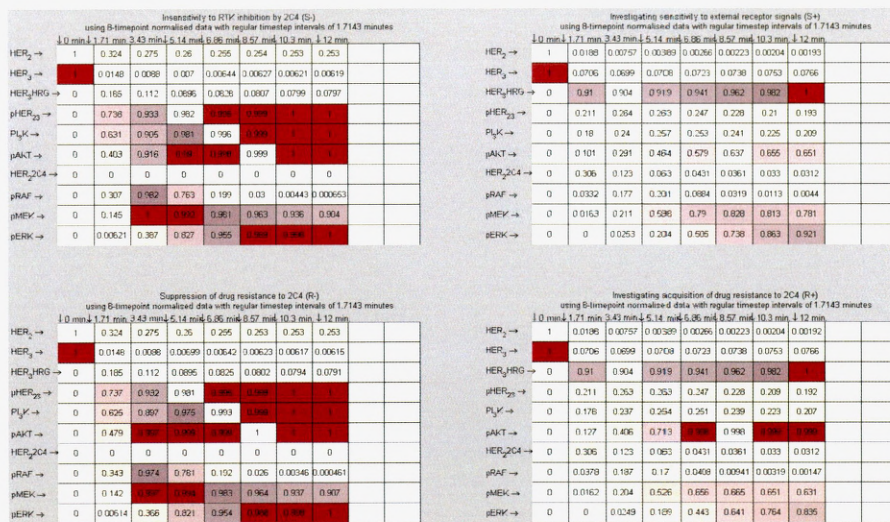


Figure D-3: Heat map representations of 4 different normalised data sets generated from a biologically plausible process-based model of HER2/3-PI3K-MAPK signalling pathways with 8-timepoint readings recorded over a period of 12 minutes only.

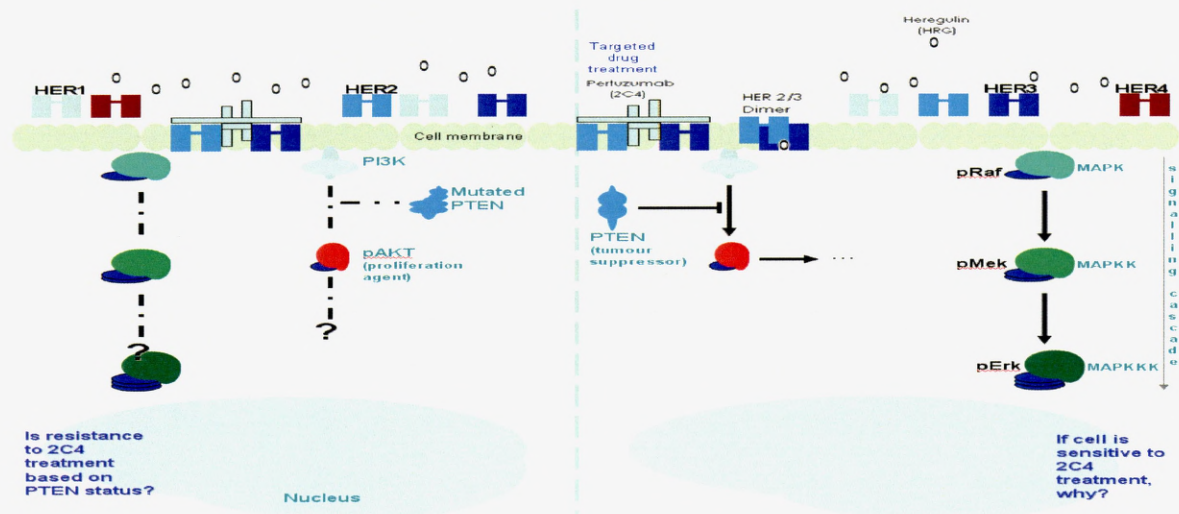


Figure D-4: Using signal transduction network data to infer or predict signalling from enzyme-coupled cell-surface receptors to intracellular kinases: modelling a) drug resistance (L.H.S) and b) sensitivity to RTK inhibition (R.H.S) purely from data generated from well-tested process-based models that switched between these two modes.

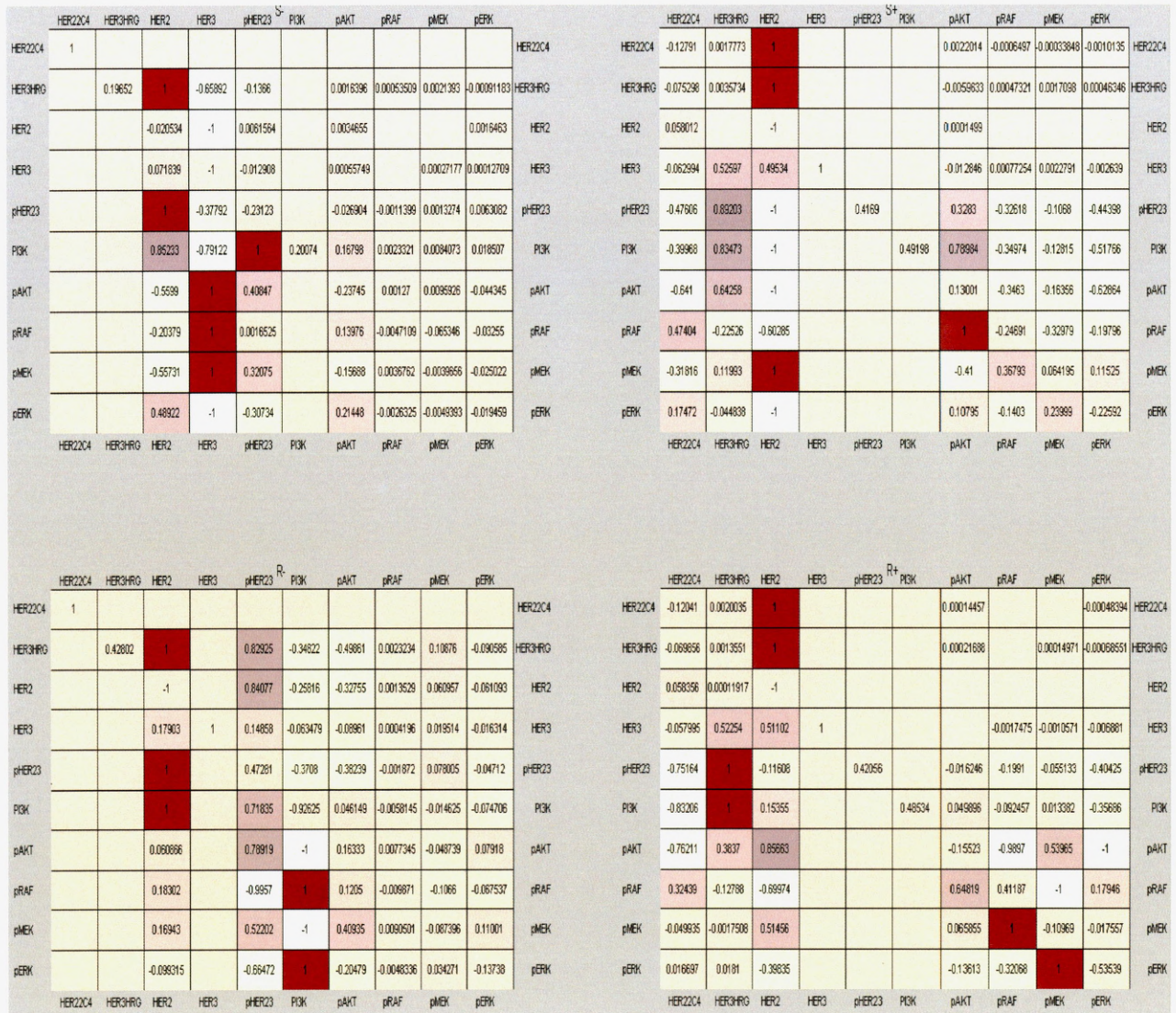


Figure D-5: Derived heatmap of signalling network of HER 2/3-MAPK/PI3K signalling pathways obtained through network inference method applied on normalised data (with only 8 time points).

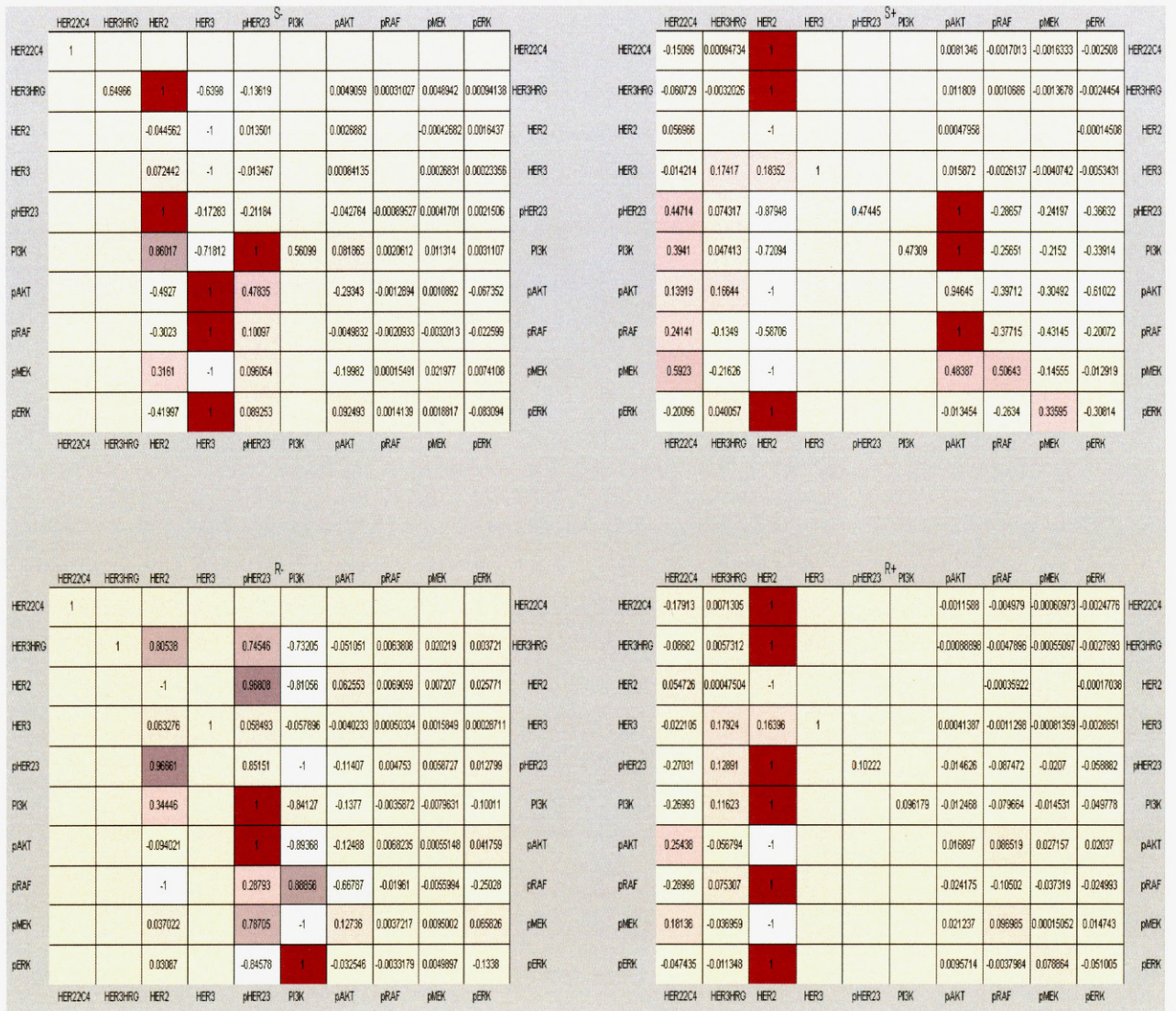


Figure D-6: Derived heatmap of signalling network of HER 2/3-MAPK/PI3K signalling pathways obtained through network inference method applied on normalised data (with only 10 time points).



Figure D-7: Derived heatmap of signalling network of HER 2/3-MAPK/PI3K signalling pathways obtained through network inference method applied on normalised data (with only 12 time points).

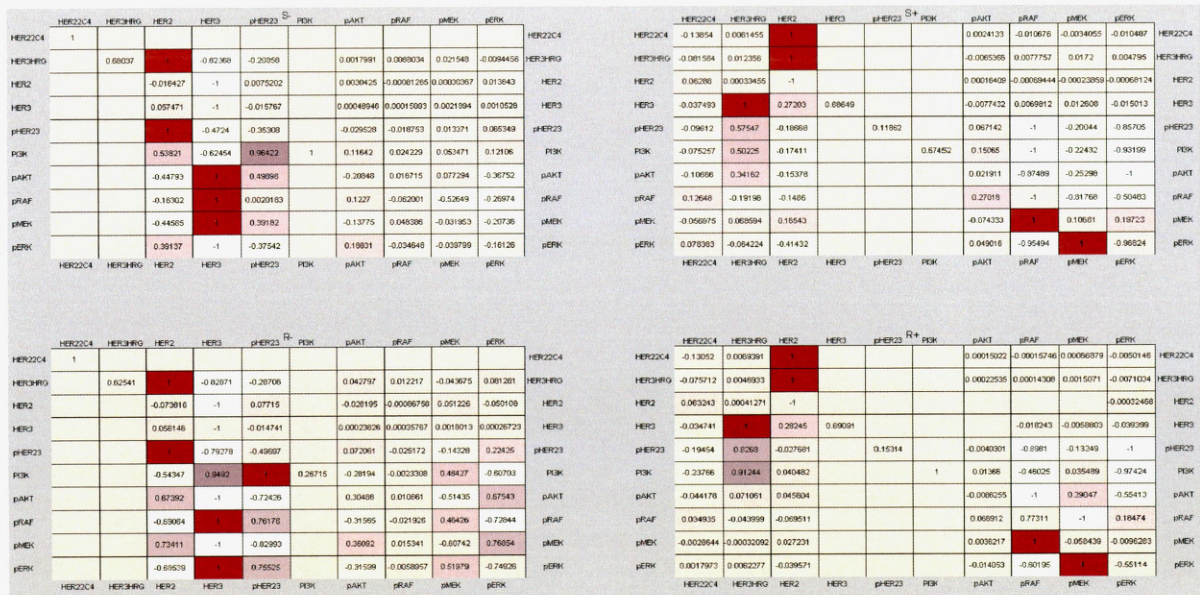


Figure D-8: Corresponding result of S-, S+, R-, R+ data, absolute data, 8 minutes.

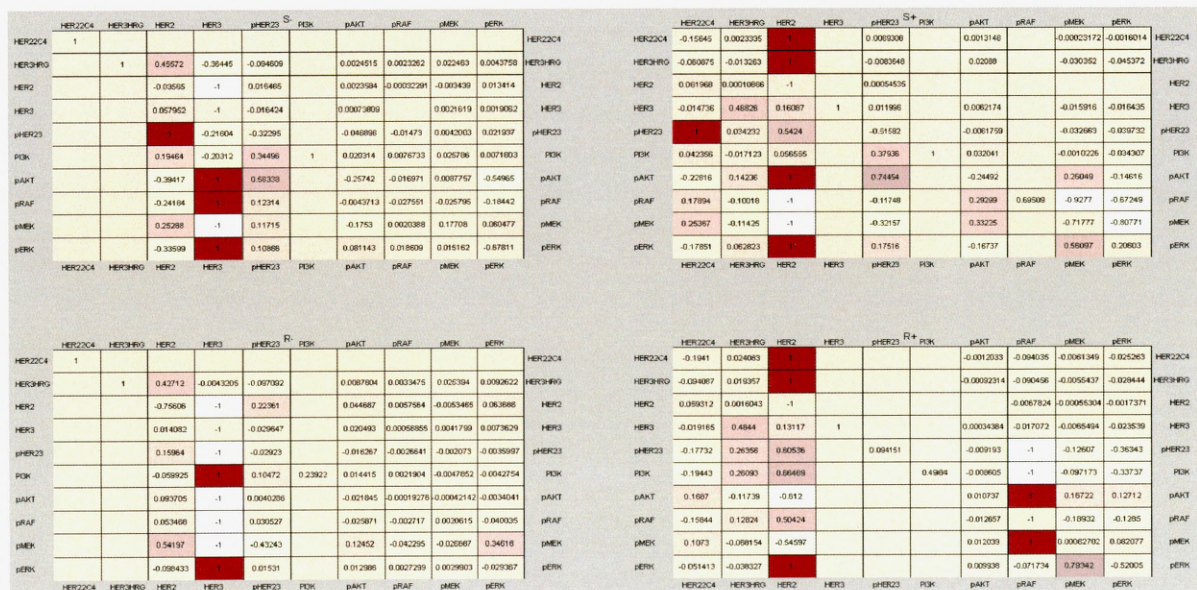


Figure D-9: Corresponding result of S-, S+, R-, R+ data, absolute data, 10 minutes.

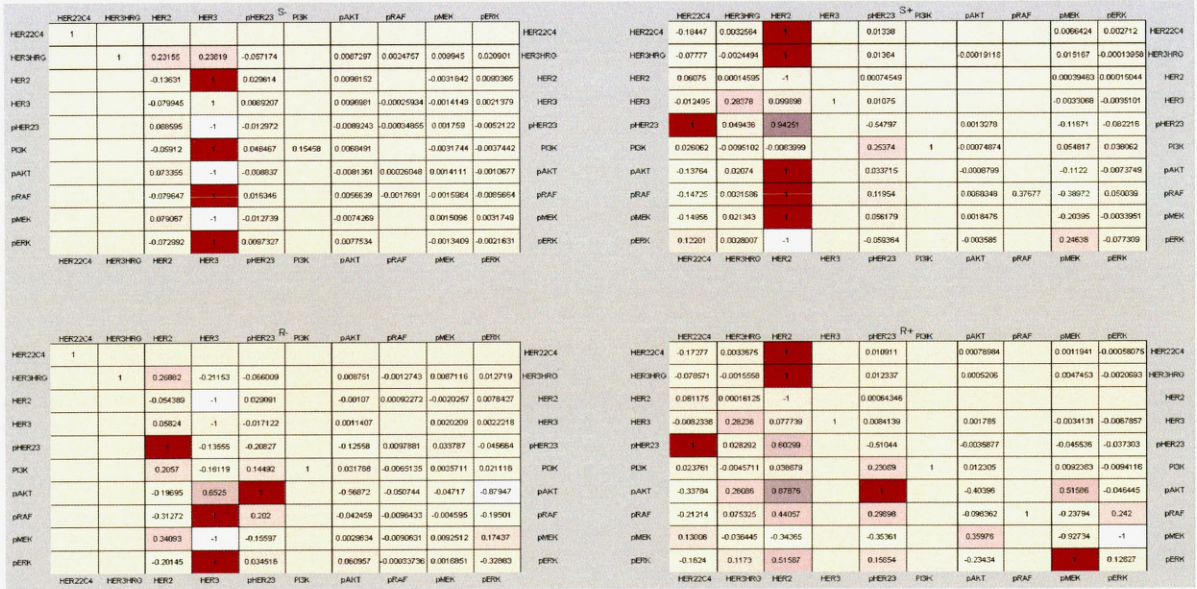


Figure D-10: Corresponding result of S-, S+, R-, R+ data, absolute data, 12 minutes.

Appendix E

Supplementary information

Using one of the sample data in chapter 6, we demonstrate how the network model of such data may be constructed, all other construction is constructed in similar fashion.

Let us say the following time series data were to processed to infer a network model:

$$\begin{bmatrix}
 time(mins.) \\
 Her_2 \\
 Her_3 \\
 Her_3Hrg \\
 pHer_{23} \\
 PI3K \\
 pAkt \\
 Her_22C_4 \\
 pRaf \\
 pMek \\
 pErk
 \end{bmatrix}
 =
 \begin{bmatrix}
 0 & \frac{8}{7} & \frac{16}{7} & \frac{24}{7} & \frac{32}{7} & \frac{40}{7} & \frac{48}{7} & 8 \\
 1 & 0.35681 & 0.30083 & 0.27459 & 0.26289 & 0.25763 & 0.2552 & 0.25407 \\
 1 & 0.018774 & 0.011996 & 0.0088041 & 0.0073793 & 0.0067392 & 0.0064443 & 0.006306 \\
 0 & 0.24551 & 0.158 & 0.11701 & 0.098736 & 0.090516 & 0.086723 & 0.084944 \\
 0 & 0.5855 & 0.83651 & 0.93496 & 0.97361 & 0.98976 & 0.99688 & 1 \\
 0 & 0.4834 & 0.75191 & 0.9066 & 0.96887 & 0.98964 & 0.99707 & 1 \\
 0 & 0.19463 & 0.62196 & 0.91637 & 0.98362 & 0.99516 & 0.99858 & 1 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0.083794 & 0.60699 & 0.98216 & 0.90939 & 0.57473 & 0.19932 & 0.056605 \\
 0 & 0.023066 & 0.43632 & 0.99983 & 0.99529 & 0.98892 & 0.98115 & 0.97012 \\
 0 & 0.00015359 & 0.049494 & 0.39413 & 0.74423 & 0.90638 & 0.97282 & 1
 \end{bmatrix}$$

Then we define $X_{(before)} =$

$$\begin{bmatrix} 1 & 0.35681 & 0.30083 & 0.27459 & 0.26289 & 0.25763 & 0.2552 \\ 1 & 0.018774 & 0.011996 & 0.0088041 & 0.0073793 & 0.0067392 & 0.0064443 \\ 0 & 0.24551 & 0.158 & 0.11701 & 0.098736 & 0.090516 & 0.086723 \\ 0 & 0.5855 & 0.83651 & 0.93496 & 0.97361 & 0.98976 & 0.99688 \\ 0 & 0.4834 & 0.75191 & 0.9066 & 0.96887 & 0.98964 & 0.99707 \\ 0 & 0.19463 & 0.62196 & 0.91637 & 0.98362 & 0.99516 & 0.99858 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.083794 & 0.60699 & 0.98216 & 0.90939 & 0.57473 & 0.19932 \\ 0 & 0.023066 & 0.43632 & 0.99983 & 0.99529 & 0.98892 & 0.98115 \\ 0 & 0.00015359 & 0.049494 & 0.39413 & 0.74423 & 0.90638 & 0.97282 \end{bmatrix}$$

,

$X_{(after)} =$

$$\begin{bmatrix} 0.35681 & 0.30083 & 0.27459 & 0.26289 & 0.25763 & 0.2552 & 0.25407 \\ 0.018774 & 0.011996 & 0.0088041 & 0.0073793 & 0.0067392 & 0.0064443 & 0.006306 \\ 0.24551 & 0.158 & 0.11701 & 0.098736 & 0.090516 & 0.086723 & 0.084944 \\ 0.5855 & 0.83651 & 0.93496 & 0.97361 & 0.98976 & 0.99688 & 1 \\ 0.4834 & 0.75191 & 0.9066 & 0.96887 & 0.98964 & 0.99707 & 1 \\ 0.19463 & 0.62196 & 0.91637 & 0.98362 & 0.99516 & 0.99858 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.083794 & 0.60699 & 0.98216 & 0.90939 & 0.57473 & 0.19932 & 0.056605 \\ 0.023066 & 0.43632 & 0.99983 & 0.99529 & 0.98892 & 0.98115 & 0.97012 \\ 0.00015359 & 0.049494 & 0.39413 & 0.74423 & 0.90638 & 0.97282 & 1 \end{bmatrix}$$

and $t_c = \frac{8}{7} - 0 = \frac{16}{7} - \frac{8}{7}$; regular time periods.

E.0.1 Calculate E_1

Use the Transposive (or Repressive) regression method (Idowu M.A. 2011b) to analyse the time series to calculate E_1 :

$$E_1 = X_{(after)} * X_{(before)}^T * ([X_{(before)} * X_{(before)}^T]^{-1})^T$$

The following result is produced:

$$real(E_1) = \begin{bmatrix} 0.78 & -0.423 & 0 & 0.0512 & 0 & 0.00287 & 0 & -0.000154 & 0.00093 & 0.00308 \\ 0.042 & -0.0233 & 0 & -0.00451 & 0 & 0.000365 & 0 & 0.0000184 & -0.0000643 & -0.0000881 \\ 0.549 & -0.303 & 0 & -0.0564 & 0 & 0.0049 & 0 & 0.000235 & -0.000797 & -0.00105 \\ 1.28 & -0.691 & 0 & 0.68 & 0 & -0.021 & 0 & -0.000858 & 0.00774 & 0.0149 \\ 0.902 & -0.418 & 0 & 0.735 & 0 & 0.0404 & 0 & 0.000828 & 0.000645 & -0.000898 \\ -0.338 & 0.532 & 0 & 1.27 & 0 & -0.0695 & 0 & 0.00368 & -0.0192 & -0.103 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 17.7 & -17.6 & 0 & -14.0 & 0 & 14.9 & 0 & -0.0983 & -3.57 & -1.82 \\ -3.55 & 3.57 & 0 & 3.05 & 0 & -0.79 & 0 & 0.0499 & -0.124 & -0.299 \\ -0.55 & 0.55 & 0 & 0.249 & 0 & 0.447 & 0 & -0.00375 & 0.12 & 0.334 \end{bmatrix}$$

Calculate J_1

Next construct the logarithmic inverse as discussed in section 4.2.8. This process is used to reverse engineer J_1 from E_1 . The following result is produced:

$$real(J_1) \approx \begin{bmatrix} -0.0629 & -3.07 & 0 & 0.0189 & -0 & 0.0106 & -0 & -0.000189 & 0.000116 & 0.00505 \\ 0.428 & -5.96 & 0 & -0.0769 & -0 & 0.00332 & -0 & 0.0000683 & 0.00162 & 0.000757 \\ 6.05 & 306.0 & -30.0 & -1.04 & 0.0267 & 0.0568 & -0.0431 & 0.00091 & 0.0213 & 0.0184 \\ 1.56 & -0.588 & -0 & -0.36 & 0 & -0.0418 & 0 & -0.00177 & 0.00206 & 0.00981 \\ -92.9 & 1444.0 & -0.532 & 32.9 & -29.4 & 8.58 & 0.0597 & 0.0602 & -0.116 & 2.52 \\ -5.54 & 9.89 & -0 & 4.04 & 0 & -2.35 & 0 & 0.0126 & 0.0949 & -0.439 \\ -0.00378 & 0.0193 & 0.00281 & 0.00117 & -0.000911 & 0.000267 & -35.6 & 0.00000192 & -0.00000652 & 0.000081 \\ -42.4 & 208.0 & -0 & 0.344 & 0 & 29.1 & 0 & -0.981 & -13.6 & -6.78 \\ -35.6 & 63.8 & -0 & 20.5 & 0 & -10.0 & 0 & 0.235 & -0.253 & -1.6 \\ 9.42 & -19.3 & 0 & -5.92 & -0 & 4.13 & -0 & -0.0507 & -0.0951 & -0.375 \end{bmatrix}$$

E.0.2 Calculating EE_1 : an alternative method

It would be reassuring to use another technique to confirm the last result. To do this, we first eliminate the 7th-row from the time series data. The technique we are about to describe may be applicable to any data with zero-row(s) or constant-row(s). We

define a constant-row as any row with a sequence of constant value in the series data, e.g. the 7th row being referred to here. This means that the states matrices $X_{(before)}$ and $X_{(after)}$ must be redefined thus:

$$X_{(before)} = \begin{bmatrix} 1 & 0.35681 & 0.30083 & 0.27459 & 0.26289 & 0.25763 & 0.2552 \\ 1 & 0.018774 & 0.011996 & 0.0088041 & 0.0073793 & 0.0067392 & 0.0064443 \\ 0 & 0.24551 & 0.158 & 0.11701 & 0.098736 & 0.090516 & 0.086723 \\ 0 & 0.5855 & 0.83651 & 0.93496 & 0.97361 & 0.98976 & 0.99688 \\ 0 & 0.4834 & 0.75191 & 0.9066 & 0.96887 & 0.98964 & 0.99707 \\ 0 & 0.19463 & 0.62196 & 0.91637 & 0.98362 & 0.99516 & 0.99858 \\ 0 & 0.083794 & 0.60699 & 0.98216 & 0.90939 & 0.57473 & 0.19932 \\ 0 & 0.023066 & 0.43632 & 0.99983 & 0.99529 & 0.98892 & 0.98115 \\ 0 & 0.00015359 & 0.049494 & 0.39413 & 0.74423 & 0.90638 & 0.97282 \end{bmatrix},$$

$$X_{(after)} = \begin{bmatrix} 0.35681 & 0.30083 & 0.27459 & 0.26289 & 0.25763 & 0.2552 & 0.25407 \\ 0.018774 & 0.011996 & 0.0088041 & 0.0073793 & 0.0067392 & 0.0064443 & 0.006306 \\ 0.24551 & 0.158 & 0.11701 & 0.098736 & 0.090516 & 0.086723 & 0.084944 \\ 0.5855 & 0.83651 & 0.93496 & 0.97361 & 0.98976 & 0.99688 & 1 \\ 0.4834 & 0.75191 & 0.9066 & 0.96887 & 0.98964 & 0.99707 & 1 \\ 0.19463 & 0.62196 & 0.91637 & 0.98362 & 0.99516 & 0.99858 & 1 \\ 0.083794 & 0.60699 & 0.98216 & 0.90939 & 0.57473 & 0.19932 & 0.056605 \\ 0.023066 & 0.43632 & 0.99983 & 0.99529 & 0.98892 & 0.98115 & 0.97012 \\ 0.00015359 & 0.049494 & 0.39413 & 0.74423 & 0.90638 & 0.97282 & 1 \end{bmatrix}$$

Again we may employ either the Transposive or Repressive) regression method (Idowu M.A. 2011b) to analyse this extracted data to construct a new matrix, EE_1 . The following step is appropriate for the reconstruction:

$$EE_1 = X_{(after)} * X_{(before)}^T * ([X_{(before)} * X_{(before)}^T]^{-1})^T$$

$$real(EE_1) =$$

$$\begin{bmatrix} 0.78 & -0.423 & 0 & 0.0512 & 0 & 0.00287 & -0.000154 & 0.00093 & 0.00308 \\ 0.042 & -0.0233 & 0 & -0.00451 & 0 & 0.000365 & 0.0000184 & -0.0000643 & -0.0000881 \\ 0.549 & -0.303 & 0 & -0.0564 & 0 & 0.0049 & 0.000235 & -0.000797 & -0.00105 \\ 1.28 & -0.691 & 0 & 0.68 & 0 & -0.021 & -0.000858 & 0.00774 & 0.0149 \\ 0.902 & -0.418 & 0 & 0.735 & 0 & 0.0404 & 0.000828 & 0.000645 & -0.000898 \\ -0.338 & 0.532 & 0 & 1.27 & 0 & -0.0695 & 0.00368 & -0.0192 & -0.103 \\ 17.7 & -17.6 & 0 & -14.0 & 0 & 14.9 & -0.0983 & -3.57 & -1.82 \\ -3.55 & 3.57 & 0 & 3.05 & 0 & -0.79 & 0.0499 & -0.124 & -0.299 \\ -0.55 & 0.55 & 0 & 0.249 & 0 & 0.447 & -0.00375 & 0.12 & 0.334 \end{bmatrix}$$

Notice that EE_1 should be identical (or almost identical) to the earlier result E_1 if E_1 's 7th-row and 7th-column were eliminated. So we may redefine EE_1 as a *reduced form* of E_1 . Next we construct the appropriate JJ_1 from EE_1 using the logarithmic inverse method in the same way J_1 was constructed in the previous section using EE_1 , the *reduced form* of E_1 , instead.

Calculating JJ_1

To *reverse engineer* JJ_1 from EE_1 the following result is produced

$real(JJ_1) \approx$

$$\begin{bmatrix} -0.0629 & -3.07 & 0 & 0.0189 & 0 & 0.0106 & -0.000189 & 0.000116 & 0.00505 \\ 0.428 & -5.96 & 0 & -0.0769 & 06 & 0.00332 & 0.0000683 & 0.00162 & 0.000757 \\ 7.51 & 295.0 & -30.8 & -1.49 & 0.377 & -0.0459 & 0.000171 & 0.0238 & -0.0129 \\ 1.56 & -0.588 & -0 & -0.36 & -0 & -0.0418 & -0.00177 & 0.00206 & 0.00981 \\ -89.7 & 1388.0 & 0.855 & 31.9 & -28.6 & 8.34 & 0.0585 & -0.111 & 2.45 \\ -5.54 & 9.89 & -0 & 4.04 & -0 & -2.35 & 0.0126 & 0.0949 & -0.439 \\ -42.4 & 208.0 & -0 & 0.344 & -0 & 29.1 & -0.981 & -13.6 & -6.78 \\ -35.6 & 63.8 & -0 & 20.5 & -0 & -10.0 & 0.235 & -0.253 & -1.6 \\ 9.42 & -19.3 & 0 & -5.92 & 0 & 4.13 & -0.0507 & -0.0951 & -0.375 \end{bmatrix}.$$

Now compare this new result JJ_1 with J_1 obtained earlier, notice the similarities and differences. Substantial discrepancies occur in rows 3,5, and 7 (though row 7 is not included in JJ_1). This suggests that appropriate handling of zero-data (or constant-

data) together with their corresponding resultant rows and columns is important and may be key to effective development of inference algorithm. Apparently the entries of J_1 and JJ_1 seem invalid (i.e. may not be data-consistent) if plugged into the ODE model with jacobian definition, i.e. both J_1 and JJ_1 have at least one parameter greater than 1300 in magnitude. Though these entries may appear unusual, the following test results demonstrate that the models still remain data-consistent.

E.0.3 Assessment of initial results

We check to see if an ODE model with the jacobian J_1 is data consistent. Such model should have the solution $X_{i+1} = \exp^{J_1 * t_c} * X_i$, where X_i is the known state before transformation, X_{i+1} is the new state after transformation, t_c (earlier defined to be equal to $\frac{8}{7}$) is the period between any two successive states, $i = 0, 1, 2, \dots$ such that X_0 is the initial condition. *Given that the initial condition X_0 is $\begin{bmatrix} 1.0 & 1.0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T$, what is the system state after $\frac{8}{7}$ minutes?* $X_1 = \exp^{(J_1 * \frac{8}{7})} * X_0$ where

$$J_1 = \begin{bmatrix} -0.0629 & -3.07 & 0 & 0.0189 & -0 & 0.0106 & -0 & -0.000189 & 0.000116 & 0.00505 \\ 0.428 & -5.96 & 0 & -0.0769 & -0 & 0.00332 & -0 & 0.0000683 & 0.00162 & 0.000757 \\ 6.05 & 306.0 & -30.0 & -1.04 & 0.0267 & 0.0568 & -0.0431 & 0.00091 & 0.0213 & 0.0184 \\ 1.56 & -0.588 & -0 & -0.36 & 0 & -0.0418 & 0 & -0.00177 & 0.00206 & 0.00981 \\ -92.9 & 1444.0 & -0.532 & 32.9 & -29.4 & 8.58 & 0.0597 & 0.0602 & -0.116 & 2.52 \\ -5.54 & 9.89 & -0 & 4.04 & 0 & -2.35 & 0 & 0.0126 & 0.0949 & -0.439 \\ -0.00378 & 0.0193 & 0.00281 & 0.00117 & -0.000911 & 0.000267 & -35.6 & 0.00 & -0.00 & 0.000081 \\ -42.4 & 208.0 & -0 & 0.344 & 0 & 29.1 & 0 & -0.981 & -13.6 & -6.78 \\ -35.6 & 63.8 & -0 & 20.5 & 0 & -10.0 & 0 & 0.235 & -0.253 & -1.6 \\ 9.42 & -19.3 & 0 & -5.92 & -0 & 4.13 & -0 & -0.0507 & -0.0951 & -0.375 \end{bmatrix}^T$$

Therefore $X_1 = \begin{bmatrix} 0.357 & 0.0188 & 0.246 & 0.585 & 0.483 & 0.195 & -0.0 & 0.0838 & 0.0231 & 0.000154 \end{bmatrix}^T$.

This result confirms that J_1 is data-consistent.

E.0.4 Major challenge

One interesting question to ponder over is how to infer the *superlative (solution) form* of J_1 and JJ_1 in a data-consistent fashion and without compromising on parameter estimates due to lack of adequate data. We use the term *superlative form* to refer to the (actual) original matrix that should be (or should have been) inferred, i.e. the most basic matrix that is associated with parameter values of small magnitudes only.

Identifying incorrect row entries (in the jacobian)

A quick-and-dirty way to identify wrong entries in the jacobian which may require further adjustment in their parameters is by comparing the result $(J_1 * X_1)$ with $(J_1 * \exp^{J_1 * \lambda} * X_1)$. Both results are expected to be equal to \dot{X} ; since $\dot{X} = J_1 * X_1 = J_1 * \exp^{J_1 * \lambda} * X_1$, where λ is an infinitesimally small number. If they are not, the rows with significant differences in values are marked for adjustment. For example, using the time series above, let $\lambda = 0.00000001$ comparing $[J_1 * X_1$ with $\exp^{(J_1 * \lambda)} * X_1]$:

$J_1 * X_1$	$J_1 * \exp^{(J_1 * \lambda)} * X_1$
-3.128	-3.128
-5.5319	-5.5319
311.77	311.77
0.9675	0.9675
1349.0	1349.0
4.3538	4.3538
0.015547	0.015547
165.75	165.75
28.237	28.237
-9.8344	-9.8344

confirms that though J_1 may be data-consistent but if $\lambda \rightarrow 0.001$ the new result

$J_1 * X_1$	$J_1 * \exp^{(J_1 * \lambda)} * X_1$	$\frac{J_1 * X_1 - J_1 * \exp^{J_1 * \lambda} * X_1}{J_1 * X_1} * 100\%$
-3.128	-3.1148	0.54578
-5.5319	-5.5076	0.5703
311.77	300.85	3.4787
0.9675	0.96556	0.2371
1349.0	1302.2	3.4594
4.3538	4.3279	0.78431
0.015547	0.014529	6.1337
165.75	164.73	0.81074
28.237	28.071	0.75916
-9.8344	-9.7736	0.81628

demonstrates that rows 3,5,7 have error differences of more 1%. To correct these weaknesses we *could continue tweaking* the rows 3,5,7 *until* all worst error differences < 1% while keeping λ constant at 0.001. One thing is certain and that is: the process of obtaining the superlative result must be standardised. Otherwise unrealistic parameter values would be unavoidable.

E.0.5 Finding a superlative (jacobian) solution

It might be necessary or required to impose some structural constraints on the inverse problem before proceeding to parameter estimation. For example, we might use J_1 to calculate:

$$\exp^{J_1 * 1} = \exp^{J_1} =$$

$$\begin{bmatrix} 0.80416 & -0.43747 & 0 & 0.045157 & 0 & 0.0032568 & 0 & -0.0001489 & 0.00073671 & 0.0028787 \\ 0.045275 & -0.02628 & 0 & -0.0052728 & 0 & 0.00029397 & 0 & 2.1673e-005 & -2.6187e-005 & -8.5209e-005 \\ 0.59053 & -0.34202 & 0 & -0.066438 & 0 & 0.0040513 & 0 & 0.00027737 & -0.00031051 & -0.00099651 \\ 1.1699 & -0.63302 & 0 & 0.70764 & 0 & -0.019608 & 0 & -0.00086809 & 0.0067155 & 0.01337 \\ 0.82048 & -0.43388 & 0 & 0.73433 & 0 & 0.057623 & 0 & 0.0010493 & 0.0017218 & 0.0017303 \\ -0.63844 & 0.82682 & 0 & 1.3438 & 0 & -0.05924 & 0 & 0.0044331 & -0.013003 & -0.11422 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 16.914 & -15.6836 & 0 & -14.2233 & 0 & 15.8615 & 0 & -0.023139 & -4.0544 & -1.8787 \\ -5.0199 & 5.3142 & 0 & 3.811 & 0 & -1.2478 & 0 & 0.060253 & -0.012113 & -0.34479 \\ -0.12232 & 0.055983 & 0 & -0.067224 & 0 & 0.5911 & 0 & -0.006607 & 0.11167 & 0.40833 \end{bmatrix}$$

which may be represented with the following adjacent matrix (network topology)

$$S = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

One superlative form of S might be the symmetrix

$$S_s = \text{booleanise}((S + S')/2)$$

such that

$$S_s = \text{booleanise} \left(\begin{bmatrix} 1 & 1 & 0.5 & 1 & 0.5 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0.5 & 1 & 0.5 & 1 & 0 & 1 & 1 & 1 \\ 0.5 & 0.5 & 0 & 0.5 & 0 & 0.5 & 0 & 0.5 & 0.5 & 0.5 \\ 1 & 1 & 0.5 & 1 & 0.5 & 1 & 0 & 1 & 1 & 1 \\ 0.5 & 0.5 & 0 & 0.5 & 0 & 0.5 & 0 & 0.5 & 0.5 & 0.5 \\ 1 & 1 & 0.5 & 1 & 0.5 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0.5 & 1 & 0.5 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0.5 & 1 & 0.5 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0.5 & 1 & 0.5 & 1 & 0 & 1 & 1 & 1 \end{bmatrix} \right)$$

$$\rightarrow S_s = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

This structure S_s can be used to constrain or redefine the reverse engineering problem before E_1 is calculated at all. Because the superlative structure S_s is slightly larger in size (in terms of the number of parameters set to 1) than the initial structure in E_1 , the new jacobian results that would be obtained by it might even be more accurate (and with parameter values of small magnitude). The parameters of the superlative structure are often more realistic than those without. Note that all diagonal entries of S_s must always be set to 1 before estimation and the flexibility to redefining the network connectivity as desired is possible.

Appendix F

Alternative methods: matrix-based analytical techniques

F.1 Method 2: heuristic development of new analytical methods

In previous works (Idowu M.A. 2011*a*, Idowu M.A. 2012) the initial development and assessment of our matrix-based network inference algorithms process was as follows: as a first step a number of promising inference methods was selected, each inference algorithm determined a set of network interactions which had to be assessed to explain the observed experimental data. That provided a baseline data set. The selected methods were developed to identify and extract only the strongly connected links within those results by eliminating all insignificant or potentially redundant associations from the initial sets of results. Subsequent resultant outcomes were then compared mathematically and computationally using well-defined network metric measures. The inference methods that successfully yielded most strongly connected links were identified and further optimised using matrix manipulation techniques including functions associated with logarithmic inverse ((Idowu M.A. 2011*a*, Idowu M.A. 2012)) and pseudoinverse ((Gilbert 1988, E.H 1920, Arne 1951, Roger 1955)) operations (Figure A-1).

Through thorough elimination of all redundancies in model reconstruction methods, significant consideration was given to understanding the essential techniques that mostly yielded positive results by comparing between the intermediate modelling results obtained from the predicted networks and those results in the hidden (expected) target network structure (Figures A-2 and A-3). No information about the original (target) network structure was supplied to any of the algorithms, i.e we ensured that the algorithms made no use of *a priori* information about network structure throughout the model reconstruction process. Setting both data consistency and network topology consistency as primary targets, it was possible to identify two fundamental inference algorithms from those sets of newly developed algorithms. Hence only those fundamental inference methods that yielded satisfactory results were chosen for further restandardisation and optimisation. Further assessment tests were performed to ensure that most of the final outcomes were both topologically and data consistent, i.e. the model structure often closely matched the original (hidden) network models that were used to simulate the conditions for the assessment tests; and the simulated time series data outputs generated from the constructed models often matched the original time series data input to the test systems as demonstrated and evidenced in (Idowu M.A. 2011a) (Figures A-4, A-5, and A-6). To further extend the work in (Idowu M.A. 2011a) we established a new reverse engineering framework (Figure 2-1) that incorporates network inference, parameter estimation, and multiple model specifications (solutions) into an integrated modelling unit. During the method refinement and restandardisation process matrix-based recast were formulated to establish new inferential procedures to further accelerate the multiple model reconstruction process, i.e. ensuring that most operations being performed are matrix-based.

F.1.1 Multiple model integration

We extend the predictive capability of discovery process by increasing the number of distinct model solutions that can be provided to an inverse problem. This is achieved by developing new methods to support jacobian to power-law model integration thereby ensuring that nonlinearities in complex systems may be captured

and modelled using ODE models designed for formulating both simple and complex nonlinear phenomena.

The novel methods presented here may be categorised into two broad groups, namely: inference or automated construction of jacobian or power-law model from time series data (straightforward or direct method); and recasting of inferred jacobian (or inferred power-law) model to construct a power-law (or jacobian) model from time series data (indirect method). The recast method, presented and discussed in the methodology section, uses a fast and powerful network inference algorithm such as the transposive regression method (TRM) presented in (Idowu M.A. 2011a) to first infer either a jacobian or power-law model from any time series data supplied before recasting it to another desirable model type or format.

F.1.2 Extending biochemical system theory (BST) framework

In BST, all system variables are represented in power-law formalism and expressions. Hence the models are said to be power-law based. Power-law based models are particularly useful for modelling dynamical systems that are often associated with high levels of non-linearity, e.g. genetic networks, metabolic networks, signal transduction network etc (Chen L. 2009).

Inverse (or system identification) problems in BST may be solved using power-law based models, e.g. half-system, S-system, or the generalised mass action (GMA) kinetic models. This usually requires employing appropriate parameter estimation methods to calibrate power-law models. Here, though S-system and GMA models are more generalised forms of BST, we focus primarily on half-system based model as a form of BST and how it may be related to a jacobian model. The quest to unveil such connection has revealed new insights into the architecture of BST and jacobian models. These special relationships provide reusable strategies for integrating jacobian models to models based on the BST framework, and vice versa.

F.2 Method 3: Half-system based inference algorithm

The half system is a form of BST which provides a complete aggregation of a system's processes to single net terms which serves as an approximation of the production (synthesis) and degradation (depletion) of the molecular constituents within the system (Voit E.O. 2000). Depending on the nature of the time series data and primary objectives of the modelling task the half-system model may be used as an effective and convenient strategy or tool for identifying and mimicking system dynamics and predicting future outcomes.

Adopting a half-system model as a nonlinear model may help ease system identification and parameter estimation challenges. This task practically involves the construction of ODE based log-linear model and applying appropriate parameter estimation techniques to infer optimal solution from time series data. For this reason, the half-system may be called a Lin-log model.

The half system representation of dynamical systems is of the form:

$$\dot{X}_i = \alpha_i \cdot \prod_{j=1}^n (X_j^{g_{ij}})$$

where $i = 1 \dots n$; n is the number of state variables; g_{ij} are called kinetic orders (models parameters) and quantify the overall net effect of each X_j on the production (or degradation) of X_i ; α_i are called rate constants. In matrix form, the half-system may be represented as:

$$\begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \\ \vdots \\ \dot{X}_n \end{bmatrix} = \begin{bmatrix} \alpha_1 \cdot \prod_{j=1}^n X_j^{g_{1j}} \\ \alpha_2 \cdot \prod_{j=1}^n X_j^{g_{2j}} \\ \vdots \\ \alpha_n \cdot \prod_{j=1}^n X_j^{g_{nj}} \end{bmatrix} \quad (\text{F.1})$$

which is equivalent to the logarithmic equations

$$\begin{bmatrix} \log(\dot{X}_1) \\ \log(\dot{X}_2) \\ \vdots \\ \log(\dot{X}_n) \end{bmatrix} = \begin{bmatrix} \log(\alpha_1) + g_{11} \log(X_1) + \dots + g_{1n} \log(X_n) \\ \log(\alpha_2) + g_{21} \log(X_1) + \dots + g_{2n} \log(X_n) \\ \vdots \\ \log(\alpha_n) + g_{n1} \log(X_1) + \dots + g_{nn} \log(X_n) \end{bmatrix} \quad (\text{F.2})$$

As a preliminary step, the logarithm of the variables and kinetic parameters and the derivatives of the variables are related:

$$\begin{bmatrix} \log(\dot{X}_1) \\ \log(\dot{X}_2) \\ \vdots \\ \log(\dot{X}_n) \end{bmatrix} = \begin{bmatrix} \log(\alpha_1) & g_{11} & g_{12} & \dots & g_{1n} \\ \log(\alpha_2) & g_{21} & g_{22} & \dots & g_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \log(\alpha_n) & g_{n1} & g_{n2} & \dots & g_{nn} \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \log(X_1) \\ \log(X_2) \\ \vdots \\ \log(X_n) \end{bmatrix} \quad (\text{F.3})$$

where all the unknown parameters are the matrix collection M

$$\begin{bmatrix} \log(\alpha_1) & g_{11} & g_{12} & \dots & g_{1n} \\ \log(\alpha_2) & g_{21} & g_{22} & \dots & g_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \log(\alpha_n) & g_{n1} & g_{n2} & \dots & g_{nn} \end{bmatrix} \quad (\text{F.4})$$

and must be inferred from the available time series data to produce a data-consistent predictive model.

The next section describes how M (the model parameters) may be estimated. The recommended solution is also based on the algorithm presented in our previous work (Idowu M.A. 2011a).

F.2.1 Half-system: estimation of kinetic parameters

The arrays of column vectors in equation F.3 required to estimate both the $\log(\alpha_i)$

and g_{ij} values in M are
$$\begin{bmatrix} \log(\dot{X}_1) \\ \log(\dot{X}_2) \\ \vdots \\ \log(\dot{X}_n) \end{bmatrix} \text{ and } \begin{bmatrix} 1 \\ \log(X_1) \\ \log(X_2) \\ \vdots \\ \log(X_n) \end{bmatrix}$$

If we let

$$X(t) = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

be a known state vector at timepoint t, then the derivative vector \dot{X} may be calculated from the two known state vectors (X_t and X_{t+1}) as $\dot{X} \approx \frac{X_{(t+1)} - X(t)}{t_c}$ where t_c is the interval of separation. The following depiction of a multi-state representation of $\log(\dot{X}(t))$ involving multiple time states

$$\begin{bmatrix} \log(\dot{X}_{10}) & \log(\dot{X}_{11}) & \dots & \log(\dot{X}_{1s-1}) \\ \log(\dot{X}_{20}) & \log(\dot{X}_{21}) & \dots & \log(\dot{X}_{2s-1}) \\ \vdots & & & \\ \log(\dot{X}_{n0}) & \log(\dot{X}_{n1}) & \dots & \log(\dot{X}_{ns-1}) \end{bmatrix} = \text{Log} \left(\begin{bmatrix} \frac{X_{11}-X_{10}}{t_c} & \frac{X_{12}-X_{11}}{t_c} & \dots & \frac{X_{1s}-X_{1s-1}}{t_c} \\ \frac{X_{21}-X_{20}}{t_c} & \frac{X_{22}-X_{21}}{t_c} & \dots & \frac{X_{2s}-X_{2s-1}}{t_c} \\ \vdots & & & \\ \frac{X_{n1}-X_{n0}}{t_c} & \frac{X_{n2}-X_{n1}}{t_c} & \dots & \frac{X_{ns}-X_{ns-1}}{t_c} \end{bmatrix} \right)$$

may be used to represent an array of column vectors $\begin{bmatrix} \log(\dot{X}_{i0}) & \log(\dot{X}_{i1}) & \dots & \log(\dot{X}_{is-1}) \end{bmatrix}$ on the L.H.S., where each k^{th} entry of $\log(\dot{X}_{it})$ is a logarithm of the derivative of the time-course data of the variable X_i at timepoint k, from its initial condition $state_0$ to state $state_{s-1}$. The following equation is implied from the previous section, which can be solved easily using one of our parameter estimation techniques, e.g. Transposive regression (Idowu M.A. 2011a). Here, we reformulate the problem into a solvable

expression.

$$\begin{bmatrix} \log(\dot{X}_{1_0}) & \log(\dot{X}_{1_1}) & \dots & \log(\dot{X}_{1_{s-1}}) \\ \log(\dot{X}_{2_0}) & \log(\dot{X}_{2_1}) & \dots & \log(\dot{X}_{2_{s-1}}) \\ \vdots & \vdots & \ddots & \vdots \\ \log(\dot{X}_{n_0}) & \log(\dot{X}_{n_1}) & \dots & \log(\dot{X}_{n_{s-1}}) \end{bmatrix} = M \cdot \begin{bmatrix} \text{ones}(1, s) \\ \log(X_{1_0}) & \log(X_{1_1}) & \dots & \log(X_{1_{s-1}}) \\ \log(X_{2_0}) & \log(X_{2_1}) & \dots & \log(X_{2_{s-1}}) \\ \vdots & \vdots & \ddots & \vdots \\ \log(X_{n_0}) & \log(X_{n_1}) & \dots & \log(X_{n_{s-1}}) \end{bmatrix}$$

This inverse problem is completely solvable using the inference method introduced in (Idowu M.A. 2011a). Given the solution the matrix M approximates all kinetic parameters g_{ij} and logarithm of the rate constants $\log(\alpha_i)$. Calculating the rate constants α_i from the first column of M is straightforward - by calculating the logarithmic exponentiation of each element of the column, since $e^{\log(\alpha_n)} = \alpha_n$ where e, the exponentiation of 1, is ≈ 2.718282 .

Next we derive new expressions that establish the relationship between a jacobian model and power-law. Ultimately, the new expressions may be used to transform a model from its jacobian form to half-system representation, and vice versa.

F.2.2 Relating the jacobian to half-system model

Half-system: reduceable parameter complexity

The Half-system

$$\left[\dot{X}_i = \alpha_i \cdot \prod_{j=1}^m (X_j^{g_{ij}}) \right]_{h_{sys}}$$

may be decomposed into two power-law products

$$\rightarrow \left[\dot{X}_i = \alpha_i \cdot \prod_{j=1}^n (X_j^{g_{ij}}) \cdot \prod_{j=n+1}^m (X_j^{g_{ij}}) \right]_{h_{sys}}$$

where n is the total number of dependent variables, $m - n$ is the total number of independent variables, i.e. number of constants. These independent variables (i.e. constants) are mathematically aggregatable, i.e. they can be combined (multiplied together) with the initial rate constants and then replaced with single new rate con-

stants, say $_{new}\alpha_i$, such that

$$_{new}\alpha_i = \alpha_i \cdot \prod_{j=n+1}^m (X_j^{g_{ij}})$$

, which is equivalent to

$$\left[\dot{X}_i = \alpha_i \cdot \prod_{j=1}^n (X_j^{g_{ij}}) * constant \right]_{h_{sys}}$$

Hence the complexity of Half-system is reduceable to the representation

$$\left[\dot{X}_i = _{new}\alpha_i \cdot \prod_{j=1}^n (X_j^{g_{ij}}) \right]_{h_{sys}}$$

since the product of all independent variables is a constant. The vector values of the dependent variables \dot{X}_i and X_i are derivable from the available time series data and we propose that the set of all kinetic parameters g_{ij} may be estimated either from the same data or indirectly from an inferred jacobian model that must be data consistent. Section F.2.1 has addressed the former, an illustration of both the former and the latter is presented in the next section.

F.2.3 Estimating fractions of kinetic parameters in pairs

Here we show for the first time how to estimate the kinetic orders (parameters) of the half-system either from the available time series data or inferred jacobian model. The method is based on the relative ratios of pairwise data of the time series. We also establish and introduce a new technique for calculating the kinetic parameters of half-system in pairs - only the ratios of pairs of kinetic parameters are calculated, the actual value of each individual parameter has to be derived from these intermediate results. We first recall the jacobian definition

$$\dot{X} = \begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \\ \vdots \\ \dot{X}_m \end{bmatrix} = J.X = \begin{bmatrix} \frac{\partial X_1}{\partial X_1} & \frac{\partial X_1}{\partial X_2} & \cdots & \frac{\partial X_1}{\partial X_m} \\ \frac{\partial X_2}{\partial X_1} & \frac{\partial X_2}{\partial X_2} & \cdots & \frac{\partial X_2}{\partial X_m} \\ \vdots & \vdots & & \vdots \\ \frac{\partial X_m}{\partial X_1} & \frac{\partial X_m}{\partial X_2} & \cdots & \frac{\partial X_m}{\partial X_m} \end{bmatrix} . X \quad (F.5)$$

and assume it is related to its equivalent half-system by the relation

$$\dot{X} = J * \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{bmatrix} = \left[\alpha_i \cdot \prod_{j=1}^m (X_j^{g_{ij}}) \right]_{h_{sys}} = \begin{bmatrix} \frac{\partial X_1}{\partial X_1} X_1 + \frac{\partial X_1}{\partial X_2} X_2 + \dots + \frac{\partial X_1}{\partial X_m} X_m \\ \frac{\partial X_2}{\partial X_1} X_1 + \frac{\partial X_2}{\partial X_2} X_2 + \dots + \frac{\partial X_2}{\partial X_m} X_m \\ \vdots \\ \frac{\partial X_m}{\partial X_1} X_1 + \frac{\partial X_m}{\partial X_2} X_2 + \dots + \frac{\partial X_m}{\partial X_m} X_m \end{bmatrix} \quad (F.6)$$

Next we observe that differentiating $\dot{X} = \left[\alpha_i \cdot \prod_{j=1}^m (X_j^{g_{ij}}) \right]_{h_{sys}}$ is further differentiated w.r.t. the dependent variable X_j produces:

$$\frac{\partial \dot{X}_i}{\partial X_j} = J_{ij} = \left[\frac{g_{ij}}{X_j} \cdot \left(\alpha_i \cdot \prod_{j=1}^m (X_j^{g_{ij}}) \right) \right] \quad (F.7)$$

In order words,

$$\begin{aligned} \frac{\partial}{\partial X_j} \left(\alpha_i \cdot \prod_{j=1}^m (X_j^{g_{ij}}(t)) \right) &= \\ \frac{g_{ij}}{X_j} \cdot \left(\alpha_i \cdot \prod_{j=1}^m (X_j^{g_{ij}}(t)) \right) &= \frac{g_{ij}}{X_j} \cdot \dot{X}_i \end{aligned} \quad (F.8)$$

This also implies that

$$J_{ij} = \frac{g_{ij}}{X_j} \cdot \dot{X}_i = \frac{g_{ij}}{X_j} \cdot \left(\frac{\partial X_i}{\partial X_1} X_1 + \frac{\partial X_i}{\partial X_2} X_2 + \dots + \frac{\partial X_i}{\partial X_m} X_m \right) \quad (F.9)$$

, i.e.

$$\dot{X}_i = \frac{X_j}{g_{ij}} \cdot \frac{\partial \dot{X}_i}{\partial X_j} \quad (F.10)$$

Similarly

$$\dot{X}_i = \frac{X_k}{g_{ik}} \cdot \frac{\partial \dot{X}_i}{\partial X_k} \quad (F.11)$$

emerges from differentiating w.r.t. the variable X_k .

Therefore

$$\dot{X}_i = \frac{X_j}{g_{ij}} \cdot \frac{\partial \dot{X}_i}{\partial X_j} = \frac{X_k}{g_{ik}} \cdot \frac{\partial \dot{X}_i}{\partial X_k} \quad (F.12)$$

It turns out that multiplying both the LHS and RHS sides by $\frac{\partial X_k}{\partial \dot{X}_i}$, i.e.

$$\dot{X}_i \cdot \frac{\partial X_k}{\partial \dot{X}_i} = \frac{X_j}{g_{ij}} \cdot \frac{\partial \dot{X}_i}{\partial X_j} \cdot \frac{\partial X_k}{\partial \dot{X}_i} = \frac{X_j}{g_{ij}} \cdot \frac{\partial X_k}{\partial X_j} = \frac{X_k}{g_{ik}} \quad (F.13)$$

and rewriting equation F.13 produces

$$\frac{\partial X_k}{\partial X_j} = \frac{g_{ij}}{X_j} \cdot \frac{X_k}{g_{ik}} = \frac{X_k}{X_j} \cdot \frac{g_{ij}}{g_{ik}} = \frac{g_{ij}}{X_j} \cdot \dot{X}_i \cdot \frac{\partial X_k}{\partial \dot{X}_i} = \frac{\partial \dot{X}_i}{\partial X_j} \cdot \frac{\partial X_k}{\partial \dot{X}_i} \quad (\text{F.14})$$

as expected. Ultimately we deduce the relation between the partial derivatives of the jacobian and the kinetic orders of the related and emergent half-system to be

$$\frac{\partial X_k}{\partial X_j} = \frac{X_k}{X_j} \cdot \frac{g_{ij}}{g_{ik}} = \frac{\left(\frac{\partial \dot{X}_i}{\partial X_j}\right)}{\left(\frac{\partial \dot{X}_i}{\partial X_k}\right)} = \frac{\text{constant}}{\text{constant}} = c \quad (\text{F.15})$$

, where c is a constant and, as demonstrated, derivable direct from the jacobian matrix. We interpret the above relation to suggest that the unknown pair of kinetic order (parameters) g_{ij} and g_{ik} to be estimated are related through the equation

$$\frac{g_{ij}}{g_{ik}} = \frac{X_j}{X_k} \cdot \frac{\left(\frac{\partial \dot{X}_i}{\partial X_j}\right)}{\left(\frac{\partial \dot{X}_i}{\partial X_k}\right)}, \quad (\text{F.16})$$

which is naturally convenient for our system identification and model reconstruction purposes because X_j and X_k can be chosen from any given time series data and the partial derivatives $\partial \dot{X}_i$ and ∂X_k are related entries of the inferred jacobian matrix, i.e. assuming the jacobian is data-consistent with the actual time series data.

F.2.4 Validating the calculated kinetic orders

To further validate the proposed method of calculating the parameters of the half-

system, we use equation F.16 which suggests $\frac{X_j}{X_k} \cdot \frac{\partial X_k}{\partial X_j} = \frac{g_{ij}}{g_{ik}} = \frac{X_j}{X_k} \cdot \frac{\left(\frac{\partial \dot{X}_i}{\partial X_j}\right)}{\left(\frac{\partial \dot{X}_i}{\partial X_k}\right)}$.

We note that if $\dot{X}_i = \left(\frac{\partial X_i}{\partial X_1} X_1 + \frac{\partial X_i}{\partial X_2} X_2 + \dots + \frac{\partial X_i}{\partial X_m} X_m\right)$ then $\frac{\partial \dot{X}_i}{\partial X_j} = \frac{\partial}{\partial X_j} \left(\frac{\partial X_i}{\partial X_1} X_1 + \frac{\partial X_i}{\partial X_2} X_2 + \dots + \frac{\partial X_i}{\partial X_m} X_m\right) = \frac{\partial}{\partial X_j} \left(\frac{\partial X_i}{\partial X_j} X_j\right)$ because the rest of the terms become zero as we differentiate w.r.t. to the variable X_j . Though the $\frac{\partial X_i}{\partial X_j}$ is a partial derivative, it is a constant number in the inferred jacobian matrix. Therefore

$$\frac{g_{ij}}{g_{ik}} = \frac{X_j}{X_k} \cdot \frac{\left(\frac{\partial \dot{X}_i}{\partial X_j}\right)}{\left(\frac{\partial \dot{X}_i}{\partial X_k}\right)} = \frac{X_j}{X_k} \cdot \frac{\left(\frac{\partial \left(\frac{\partial X_i}{\partial X_j} X_j\right)}{\partial X_j}\right)}{\left(\frac{\partial \left(\frac{\partial X_i}{\partial X_k} X_k\right)}{\partial X_k}\right)} = \frac{X_j}{X_k} \cdot \frac{\left(\frac{\partial X_i}{\partial X_j}\right)}{\left(\frac{\partial X_i}{\partial X_k}\right)} \quad (\text{F.17})$$

validates that

$$\frac{g_{ij}}{g_{ik}} = \frac{X_j}{X_k} \cdot \frac{\left(\frac{\partial X_i}{\partial X_j}\right)}{\left(\frac{\partial X_i}{\partial X_k}\right)} = \frac{X_j}{X_k} \cdot \frac{J_{ij}}{J_{ik}} \quad (\text{F.18})$$

must be true.

F.2.5 Vectorisation of estimated ratios of kinetic orders (parameters)

The estimated ratios of the kinetic orders of the half-system may need to be assembled in parallel (i.e. reformulated in vectorised format) to accelerate the estimation process. The following illustrates this:

$$\begin{bmatrix} \frac{g_{ii}}{g_{i1}} & \frac{g_{ii}}{g_{i2}} & \dots & \frac{g_{ii}}{g_{im}} \end{bmatrix} = \begin{bmatrix} \frac{X_i}{X_1} \cdot \frac{\frac{\partial X_i}{\partial X_1}}{\frac{\partial X_i}{\partial X_1}} & \frac{X_i}{X_2} \cdot \frac{\frac{\partial X_i}{\partial X_1}}{\frac{\partial X_i}{\partial X_2}} & \dots & \frac{X_i}{X_m} \cdot \frac{\frac{\partial X_i}{\partial X_1}}{\frac{\partial X_i}{\partial X_m}} \end{bmatrix} \quad (\text{F.19})$$

Working with vectors of parameters is much better than estimating individual kinetic parameter one after the other.

F.2.6 Matriculation of estimated ratios of kinetic orders (parameters)

Similarly the estimated ratios of the kinetic orders may be further rearranged in matrix form to further speed up the estimation process. Extending the last expression to build a matrix of row vectors in terms of ratios of the kinetic parameters yields:

$$\begin{bmatrix} \frac{g_{11}}{g_{11}} & \frac{g_{11}}{g_{12}} & \dots & \frac{g_{11}}{g_{1m}} \\ \frac{g_{22}}{g_{21}} & \frac{g_{22}}{g_{22}} & \dots & \frac{g_{22}}{g_{2m}} \\ \vdots & \vdots & & \vdots \\ \frac{g_{mm}}{g_{m1}} & \frac{g_{mm}}{g_{m2}} & \dots & \frac{g_{mm}}{g_{mm}} \end{bmatrix} = \begin{bmatrix} \frac{X_1}{X_1} \cdot \frac{\frac{\partial X_1}{\partial X_1}}{\frac{\partial X_1}{\partial X_1}} & \frac{X_1}{X_2} \cdot \frac{\frac{\partial X_1}{\partial X_1}}{\frac{\partial X_1}{\partial X_2}} & \dots & \frac{X_1}{X_m} \cdot \frac{\frac{\partial X_1}{\partial X_1}}{\frac{\partial X_1}{\partial X_m}} \\ \frac{X_2}{X_1} \cdot \frac{\frac{\partial X_2}{\partial X_1}}{\frac{\partial X_2}{\partial X_1}} & \frac{X_2}{X_2} \cdot \frac{\frac{\partial X_2}{\partial X_1}}{\frac{\partial X_2}{\partial X_2}} & \dots & \frac{X_2}{X_m} \cdot \frac{\frac{\partial X_2}{\partial X_1}}{\frac{\partial X_2}{\partial X_m}} \\ \vdots & \vdots & & \vdots \\ \frac{X_m}{X_1} \cdot \frac{\frac{\partial X_m}{\partial X_1}}{\frac{\partial X_m}{\partial X_1}} & \frac{X_m}{X_2} \cdot \frac{\frac{\partial X_m}{\partial X_1}}{\frac{\partial X_m}{\partial X_2}} & \dots & \frac{X_m}{X_m} \cdot \frac{\frac{\partial X_m}{\partial X_1}}{\frac{\partial X_m}{\partial X_m}} \end{bmatrix} \quad (\text{F.20})$$

which can be decomposed into the following expression:

$$\begin{bmatrix} g_{11} & 0 \dots 0 \\ 0 & g_{22} \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots g_{mm} \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{g_{11}} & \frac{1}{g_{12}} & \dots & \frac{1}{g_{1m}} \\ \frac{1}{g_{21}} & \frac{1}{g_{22}} & \dots & \frac{1}{g_{2m}} \\ \vdots & \vdots & & \vdots \\ \frac{1}{g_{m1}} & \frac{1}{g_{m2}} & \dots & \frac{1}{g_{mm}} \end{bmatrix} = \begin{bmatrix} 1 & \frac{X_1}{X_2} \cdot \frac{\partial X_1}{\partial X_2} & \dots & \frac{X_1}{X_m} \cdot \frac{\partial X_1}{\partial X_m} \\ \frac{X_2}{X_1} \cdot \frac{\partial X_2}{\partial X_1} & 1 & \dots & \frac{X_2}{X_m} \cdot \frac{\partial X_2}{\partial X_m} \\ \vdots & \vdots & & \vdots \\ \frac{X_m}{X_1} \cdot \frac{\partial X_m}{\partial X_1} & \frac{X_m}{X_2} \cdot \frac{\partial X_m}{\partial X_2} & \dots & 1 \end{bmatrix} \quad (\text{F.21})$$

An alternative representation that is also valid is:

$$\begin{bmatrix} \frac{1}{g_{11}} & 0 \dots 0 \\ 0 & \frac{1}{g_{22}} \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots \frac{1}{g_{mm}} \end{bmatrix} \cdot \begin{bmatrix} g_{11} & g_{12} \dots g_{1m} \\ g_{21} & g_{22} \dots g_{2m} \\ \vdots & \vdots \\ g_{m1} & g_{m2} \dots g_{mm} \end{bmatrix} = \begin{bmatrix} 1 & \frac{X_2}{X_1} \cdot \frac{\partial X_1}{\partial X_2} & \dots & \frac{X_m}{X_1} \cdot \frac{\partial X_1}{\partial X_m} \\ \frac{X_1}{X_2} \cdot \frac{\partial X_2}{\partial X_1} & 1 & \dots & \frac{X_m}{X_2} \cdot \frac{\partial X_2}{\partial X_m} \\ \vdots & \vdots & & \vdots \\ \frac{X_1}{X_m} \cdot \frac{\partial X_m}{\partial X_1} & \frac{X_2}{X_m} \cdot \frac{\partial X_m}{\partial X_2} & \dots & 1 \end{bmatrix} \quad (\text{F.22})$$

The following equivalent symbolic representations and identities emerge:

$$G_{diag}^{-1} * G = X_{diag}^{-1} * (J_{diag}^{-1} * J) * X_{diag}; \quad (\text{F.23})$$

$$G_{diag}^{-1} * G = (J_{diag} * X_{diag})^{-1} * (J * X_{diag}); \quad (\text{F.24})$$

$$G_{diag}^{-1} * G = (X_{diag} * J_{diag})^{-1} * J * X_{diag}. \quad (\text{F.25})$$

The last expression emerges from natural properties of diagonal matrices (Gilbert 1988).

F.2.7 Inverse diagonalisation of the principal entries of the jacobian

The (reciprocal) inverse of the principal entries of the jacobian set may be diagonalised, factorised from the derived expression

$$G_{diag}^{-1} * G = (J_{diag} * X_{diag})^{-1} * J * X_{diag}$$

as

$$G_{diag}^{-1} * G = ((J * X_{diag})_{diag})^{-1} * J * X_{diag}$$

$$G_{diag}^{-1} * G = J_{diag}^{-1} * (X_{diag}^{-1} * J * X_{diag})$$

implying that

$$G_{diag}^{-1} * G = \begin{bmatrix} \frac{1}{(\frac{\partial X_1}{\partial X_1})} & 0 \dots 0 \\ 0 & \frac{1}{(\frac{\partial X_2}{\partial X_2})} \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots \frac{1}{(\frac{\partial X_m}{\partial X_m})} \end{bmatrix} \cdot \begin{bmatrix} (\frac{\partial X_1}{\partial X_1}) & \frac{X_2}{X_1} \cdot (\frac{\partial X_1}{\partial X_2}) \dots & \frac{X_m}{X_1} \cdot (\frac{\partial X_1}{\partial X_m}) \\ \frac{X_1}{X_2} \cdot (\frac{\partial X_2}{\partial X_1}) & (\frac{\partial X_2}{\partial X_2}) \dots & \frac{X_m}{X_2} \cdot (\frac{\partial X_2}{\partial X_m}) \\ \vdots & \vdots & \vdots \\ \frac{X_1}{X_m} \cdot (\frac{\partial X_m}{\partial X_1}) & \frac{X_2}{X_m} \cdot (\frac{\partial X_m}{\partial X_2}) \dots & (\frac{\partial X_m}{\partial X_m}) \end{bmatrix} \quad (F.26)$$

, which is equivalent to the generalisation

$$G_{diag}^{-1} * G = \begin{bmatrix} \frac{1}{(\frac{\partial X_1}{\partial X_1})} & 0 \dots 0 \\ 0 & \frac{1}{(\frac{\partial X_2}{\partial X_2})} \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots \frac{1}{(\frac{\partial X_m}{\partial X_m})} \end{bmatrix} \cdot \begin{bmatrix} \frac{X_1}{X_1} \cdot (\frac{\partial X_1}{\partial X_1}) & \frac{X_2}{X_1} \cdot (\frac{\partial X_1}{\partial X_2}) \dots & \frac{X_m}{X_1} \cdot (\frac{\partial X_1}{\partial X_m}) \\ \frac{X_1}{X_2} \cdot (\frac{\partial X_2}{\partial X_1}) & \frac{X_2}{X_2} \cdot (\frac{\partial X_2}{\partial X_2}) \dots & \frac{X_m}{X_2} \cdot (\frac{\partial X_2}{\partial X_m}) \\ \vdots & \vdots & \vdots \\ \frac{X_1}{X_m} \cdot (\frac{\partial X_m}{\partial X_1}) & \frac{X_2}{X_m} \cdot (\frac{\partial X_m}{\partial X_2}) \dots & \frac{X_m}{X_m} \cdot (\frac{\partial X_m}{\partial X_m}) \end{bmatrix} \quad (F.27)$$

by nature. Since

$$G_{diag}^{-1} * G = J_{diag}^{-1} * X_{diag}^{-1} * J * X_{diag} \quad (F.28)$$

$$= \begin{bmatrix} \frac{1}{(\frac{\partial X_1}{\partial X_1})} & 0 \dots 0 \\ 0 & \frac{1}{(\frac{\partial X_2}{\partial X_2})} \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots \frac{1}{(\frac{\partial X_m}{\partial X_m})} \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{X_1} & 0 \dots 0 \\ 0 & \frac{1}{X_2} \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots \frac{1}{X_m} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial X_1}{\partial X_1} & \frac{\partial X_1}{\partial X_2} \dots & \frac{\partial X_1}{\partial X_m} \\ \frac{\partial X_2}{\partial X_1} & \frac{\partial X_2}{\partial X_2} \dots & \frac{\partial X_2}{\partial X_m} \\ \vdots & \vdots & \vdots \\ \frac{\partial X_m}{\partial X_1} & \frac{\partial X_m}{\partial X_2} \dots & \frac{\partial X_m}{\partial X_m} \end{bmatrix} \cdot \begin{bmatrix} X_1 & 0 \dots 0 \\ 0 & X_2 \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots X_m \end{bmatrix}$$

which is an elegant and sophisticated representation for the expressing the quantities of $G_{diag}^{-1} * G$.

F.2.8 Derivation of the kinetic orders matrix of the half-system model

The matrix of all kinetic orders G may be estimated with a set of diagonal entries using the derivation

$$G_{diag} * (J_{diag} * X_{diag})^{-1} * J * X_{diag}$$

or

$$G = G_{diag} * X_{diag}^{-1} * J_{diag}^{-1} * J * X_{diag}$$

which means that the target G_{diag} may be fixed or predetermined as desired yielding the following matrix form:

$$G = \begin{bmatrix} g_{11} & 0 \dots 0 \\ 0 & g_{22} \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots g_{mm} \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{X_1} & 0 \dots 0 \\ 0 & \frac{1}{X_2} \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots \frac{1}{X_m} \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{(\frac{\partial X_1}{\partial X_1})} & 0 \dots 0 \\ 0 & \frac{1}{(\frac{\partial X_2}{\partial X_2})} \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots \frac{1}{(\frac{\partial X_m}{\partial X_m})} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial X_1}{\partial X_1} & \frac{\partial X_1}{\partial X_2} \dots \frac{\partial X_1}{\partial X_m} \\ \frac{\partial X_2}{\partial X_1} & \frac{\partial X_2}{\partial X_2} \dots \frac{\partial X_2}{\partial X_m} \\ \vdots & \vdots \\ \frac{\partial X_m}{\partial X_1} & \frac{\partial X_m}{\partial X_2} \dots \frac{\partial X_m}{\partial X_m} \end{bmatrix} \cdot \begin{bmatrix} X_1 & 0 \dots 0 \\ 0 & X_2 \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots X_m \end{bmatrix} \quad (F.29)$$

¹.

F.2.9 Relation between the jacobian and kinetic orders matrix

From the expression

$$G = G_{diag} * X_{diag}^{-1} * J_{diag}^{-1} * J * X_{diag}$$

¹We may use the last expression in eq. F.29 to obtained different sets of kinetic parameters for different subsets of time series data

we may derive an expression for the jacobian model using the formula

$$J = J_{diag} * X_{diag} * G_{diag}^{-1} * G * X_{diag}^{-1}$$

which, in matrix form, is:

$$J = \begin{bmatrix} (\frac{\partial X_1}{\partial X_1}) & 0 \dots 0 \\ 0 & (\frac{\partial X_2}{\partial X_2}) \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots (\frac{\partial X_m}{\partial X_m}) \end{bmatrix} \cdot \begin{bmatrix} X_1 & 0 \dots 0 \\ 0 & X_2 \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots X_3 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{g_{11}} & 0 \dots 0 \\ 0 & \frac{1}{g_{22}} \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots \frac{1}{g_{mm}} \end{bmatrix} \cdot \begin{bmatrix} g_{11} & g_{12} \dots g_{1m} \\ g_{21} & g_{22} \dots g_{2m} \\ \vdots & \vdots \\ g_{m1} & g_{m2} \dots g_{mm} \end{bmatrix} * \begin{bmatrix} \frac{1}{X_1} & 0 \dots 0 \\ 0 & \frac{1}{X_2} \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots \frac{1}{X_3} \end{bmatrix} \quad (F.30)$$

F.2.10 Significant contribution to BST: new recast technique (BAE)

Our theoretical contribution to BST in terms of how an inferred jacobian model of a time series data of a dynamical system may be related to a derived data-consistent half-system model is summarised. The first among our discoveries is the expression coined bidirectional associativity expression (BAE). BAE establishes a direct link between the jacobian and the matrix of kinetic orders for the jacobian model and half-system model. Symbolically stated BAE is

$$X_{diag} * G_{diag}^{-1} * G = J_{diag}^{-1} * J * X_{diag} \quad (F.31)$$

which in matrix form is represented as

$$\begin{bmatrix} X_1 & 0 \dots 0 \\ 0 & X_2 \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots X_3 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{g_{11}} & 0 \dots 0 \\ 0 & \frac{1}{g_{22}} \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots \frac{1}{g_{mm}} \end{bmatrix} \cdot \begin{bmatrix} g_{11} & g_{12} \dots g_{1m} \\ g_{21} & g_{22} \dots g_{2m} \\ \vdots & \vdots \\ g_{m1} & g_{m2} \dots g_{mm} \end{bmatrix} =$$

$$\begin{bmatrix} \frac{1}{(\frac{\partial X_1}{\partial X_1})} & 0 \dots 0 \\ 0 & \frac{1}{(\frac{\partial X_2}{\partial X_2})} \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots \frac{1}{(\frac{\partial X_m}{\partial X_m})} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial X_1}{\partial X_1} & \frac{\partial X_1}{\partial X_2} \dots \frac{\partial X_1}{\partial X_m} \\ \frac{\partial X_2}{\partial X_1} & \frac{\partial X_2}{\partial X_2} \dots \frac{\partial X_2}{\partial X_m} \\ \vdots & \vdots \\ \frac{\partial X_m}{\partial X_1} & \frac{\partial X_m}{\partial X_2} \dots \frac{\partial X_m}{\partial X_m} \end{bmatrix} \cdot \begin{bmatrix} X_1 & 0 \dots 0 \\ 0 & X_2 \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots X_3 \end{bmatrix} \quad (\text{F.32})$$

such that

$$\begin{bmatrix} \log(\dot{X}_1) \\ \log(\dot{X}_2) \\ \vdots \\ \log(\dot{X}_n) \end{bmatrix} = \begin{bmatrix} \log(\alpha_1) & g_{11} & g_{12} & \dots & g_{1n} \\ \log(\alpha_2) & g_{21} & g_{22} & \dots & g_{2n} \\ & & \vdots & & \\ \log(\alpha_n) & g_{n1} & g_{n2} & \dots & g_{nn} \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \log(X_1) \\ \log(X_2) \\ \vdots \\ \log(X_n) \end{bmatrix} \quad (\text{F.33})$$

$$\begin{bmatrix} \log(\dot{X}_1) \\ \log(\dot{X}_2) \\ \vdots \\ \log(\dot{X}_n) \end{bmatrix} = \begin{bmatrix} \log(\alpha_1) \\ \log(\alpha_2) \\ \vdots \\ \log(\alpha_n) \end{bmatrix} + \begin{bmatrix} g_{11} & g_{12} & \dots & g_{1n} \\ g_{21} & g_{22} & \dots & g_{2n} \\ & \vdots & & \\ g_{n1} & g_{n2} & \dots & g_{nn} \end{bmatrix} \begin{bmatrix} \log(X_1) \\ \log(X_2) \\ \vdots \\ \log(X_n) \end{bmatrix} \quad (\text{F.34})$$

$$\begin{bmatrix} \log(\dot{X}_1) \\ \log(\dot{X}_2) \\ \vdots \\ \log(\dot{X}_n) \end{bmatrix} - \begin{bmatrix} \log(\alpha_1) \\ \log(\alpha_2) \\ \vdots \\ \log(\alpha_n) \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} & \dots & g_{1n} \\ g_{21} & g_{22} & \dots & g_{2n} \\ & \vdots & & \\ g_{n1} & g_{n2} & \dots & g_{nn} \end{bmatrix} \begin{bmatrix} \log(X_1) \\ \log(X_2) \\ \vdots \\ \log(X_n) \end{bmatrix} \quad (\text{F.35})$$

$$G_{diag} = \begin{bmatrix} g_{11} & 0 \dots 0 \\ 0 & g_{22} \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots g_{mm} \end{bmatrix} = \begin{bmatrix} \frac{J_{11} \cdot X_1}{J_1 \cdot X} & 0 \dots 0 \\ 0 & \frac{J_{22} \cdot X_2}{J_2 \cdot X} \dots 0 \\ \vdots & \vdots \\ 0 & 0 \dots \frac{J_{mm} \cdot X_m}{J_m \cdot X} \end{bmatrix} \quad (\text{F.36})$$

We note that in the BST half-system representation the rate constants α_i ($i =$

$1, 2, \dots, n$) must be consistent with the expression

$$\begin{bmatrix} \log(\alpha_1) \\ \log(\alpha_2) \\ \vdots \\ \log(\alpha_n) \end{bmatrix} = \begin{bmatrix} \log(\dot{X}_1) \\ \log(\dot{X}_2) \\ \vdots \\ \log(\dot{X}_n) \end{bmatrix} - \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ & & \ddots & \\ g_{n1} & g_{n2} & \cdots & g_{nn} \end{bmatrix} \begin{bmatrix} \log(X_1) \\ \log(X_2) \\ \vdots \\ \log(X_n) \end{bmatrix}$$

which we find to be highly convenient because

$$G.X_{log} = \begin{bmatrix} \log(\frac{\dot{X}_1}{\alpha_1}) \\ \log(\frac{\dot{X}_2}{\alpha_2}) \\ \vdots \\ \log(\frac{\dot{X}_n}{\alpha_n}) \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ & & \ddots & \\ g_{n1} & g_{n2} & \cdots & g_{nn} \end{bmatrix} \cdot \begin{bmatrix} \log(X_1) \\ \log(X_2) \\ \vdots \\ \log(X_n) \end{bmatrix} \quad (F.37)$$

$$\log(\dot{X}) - \log(\alpha) = \log(J.X) - \log(\alpha) =$$

$$G.X_{log} = G_{diag} * X_{diag}^{-1} * J_{diag}^{-1} * J * X_{diag}.X_{log} \quad (F.38)$$

$$\log(\alpha) = \log(J_{diag} * X_{diag} * G_{diag}^{-1} * G * X_{diag}^{-1}.X) -$$

$$G_{diag} * X_{diag}^{-1} * J_{diag}^{-1} * J * X_{diag}.X_{log} \quad (F.39)$$

F.2.11 Application of new recast method to real experimental data

Figures B-1 depict the set of Half-system models derived using the recast technique introduced in this section. The matrices in figure B-1 are derived using the information in figure 5-4, i.e. the matrices in figure 5-4 (chapter 4).

Figure B-1 shows the result of obtaining other solutions that are dependent of information obtained from one of the four models, for instance, we used the information about the diagonal entries in the $0.1\mu\text{M}$ Dox (top-left) result to control and influence the nature of the outcomes generated in the other three results. Though figure B-1 is not used in our final analysis (it demonstrates three main possibilities: 1) imposing structural constraints (i.e. information about part of the network) of one network on

others to steer and determine their final outcomes, whenever necessary; ii) *enabling* multiple models inferred from different time series data sets to be *fixated* and made comparable to each other - this is useful when comparative study is required; and iii) aligning all inferred models to common, dependent, and shared scaling factors. Notice that in figure B-1 the level of (colour) contrast is on average less than that in figure 5-4. This is because in figure B-1 common constraints had been imposed to steer the inferred outcomes. This steering was done by fixing all the diagonals of the matrices of the kinetic orders (parameters) to the same value. It may be necessary sometimes to keep all the diagonals of the matrices of multiple models constant to measure how varied the non-diagonal parameters are approximated across the models. Comparing the magnitudes of these non-diagonal parameters and their perturbation profiles may give new insights into the topologies of the systems from a perspective different from normal network inference. These perturbation profiles (i.e. fluctuation or difference in parameters) may be evidenced by the colour contrast displayed in the generated heatmaps of the inferred models.

F.2.12 Conclusion: power-law method

Ordinary differential equations (ODEs) are commonly used to describe dynamical systems. Whenever time series profiles of constituents of a complex dynamical system become available, such time-evolution dynamics may be described either by a set of ODEs, e.g. jacobian or power-law model. Such time series evolution are described in mathematical terms that capture nonlinear dynamics of system behaviour and states recorded at various time points and intervals.

One of the most difficult challenges in modelling biological systems from time series data is the determination of a data-consistent solution to its model reconstruction or system identification problem, i.e. inverse problems. Solving an inverse problem often requires developing or appropriating effective system identification and parameter estimation strategies to (re)construct a predictive model calibrated with optimal set of parameters in a workable and data consistent way. Depending on the nature of the systems, the modeller may adopt a power-law model, either as a complementary

approach to other existing approaches or as an alternative means of formulating and validating system behaviours or dynamics through modelling. In formulating and describing complex biological processes, it is important to take into account a consideration of the underlying nonlinear phenomena involved and all the essential relationships among the system components.

A half-system is a form of BST (power-law) based model which can be used to approximate and articulate highly complex system dynamics in meaningful ways. The recast technique introduced and presented here is an integral part of a new theoretical framework that is currently being developed to extend the capabilities of both our reverse engineering and current BST frameworks to support automated time series data modelling in systems biology and beyond. The modelling approach we have described is extremely fast, optimised, and completely data-driven. The method is applicable to any time series data of dynamical systems, i.e. with unknown underlying network of interactions. Multiple data-consistent models may be inferred from such time series data without requiring *a priori* information about the architecture of the target systems.

Appendix G

New matrix construction and decomposition methods

We present a new method for constructing and decomposing square matrices. This method, based on the computed parameterisation of their implied determinants and minors, operates on the product of factors of a new form of matrix decomposition. This method may be employed to build new matrices with fixed determinant(s). We demonstrate that this new approach is fundamentally well-connected to the Cholesky decomposition if applied on symmetric matrices. We also demonstrate that it is related to the LU decomposition method via a diagonal matrix multiplier. Also through this new method a direct relation between Cholesky decomposition and LU factorisation is shown. This method, presented for the first time, is useful for (re)constructing matrices with a predefined determinant and simulating inverse problems. The inference method introduced here also is based on new matrix manipulation techniques that we have developed for the identification of systems from reproducible time series data.

In systems biology, where theoretical models are important aids in interpreting complex systems dynamics, a robust framework that is inexpensive and able to simplify the creation and evaluation of system identification and parameter estimation problems and solutions is valuable. The framework we have developed is matrix-based and sophisticated enough for the identification of ODE models from time series

data. We demonstrate that through simple matrix manipulation techniques, powerful and effective computational tools, complementary to existing reverse engineering and modelling packages, may be developed. These techniques are useful for understanding complex network structures and dynamics. We present a new method for constructing and decomposing square matrices. This method, based on the computed parameterisation of their implied determinants and minors, operates on the product of factors of a new form of matrix decomposition and may be employed to build new matrices with fixed determinant(s).

Suppose the square matrix A is to be partitioned into LDU decomposition factors, i.e., A is to be transformed into equivalent lower, diagonal, and upper matrix factors, then one may employ the Gaussian elimination method (Gilbert 1988). Importantly, here we show that A may be decomposed and recomposed in terms of its implied determinant and minors. Given A may be factorised in terms of its minor and determinants, we examine the determinant of A we suggest that if the entries of the LDU factors of A are modified such that the determinant value remains fixed, then any matrix that is reconstructed as a product of those modified LDU factors will have its determinant equal to that fixed value. Thus we propose that by this new method multiple nonsingular matrices with a predefined determinant may be created. This algorithm, presented here for the first time, provides a robust method for constructing nonsingular or singular matrix with a predefined determinant and is applicable to matrices of different sizes. This method provides us with a new tool for decomposing square matrices. We demonstrate that for symmetric matrices our new LDU decomposition method is related to the Cholesky decomposition method (Gilbert 1988) and to the LU decomposition method (Gilbert 1988).

G.1 A new matrix decomposition and composition method

It was Householder (A.S 1975) who first hinted that when an LDU factorization exists and is unique there might be a closed (explicit) formula for the elements of the L, D, and U factors in terms of the ratios of the determinants of certain submatrices of the original matrix A. However, Householder did not explain how to determine this.

Matrix decomposition is used in matrix algebra to solve systems of linear equations. An LDU factorisation of a matrix, A, is basically a decomposition of the matrix to the form $A = L.D.U$ where L and U are lower- and upper-unit triangular matrices, respectively, and D is a diagonal matrix. To develop a method for constructing matrices with a predefined determinant, we discovered that an optimised variant of the LDU technique is necessary. This decomposition technique, initially viewed as an optimised reverse engineering method of matrix composition, may be viewed from the perspective that the product of $L_d.D_d.U_d$ factors is also useful for generating reproducible time series data. Expressed in terms of the parameterisation of its implied minors and determinant, it can be used to create nonsingular jacobian matrices. As demonstrated later, the L_d and U_d factors are triangular matrices but not necessarily unit-triangular. From the definition $A = L_d.D_d.U_d$, reading from LHS to RHS, we may view the decomposition process as transforming the matrix A to $L_d.D_d.U_d$, while on the other hand, interpreting from RHS to LHS, a composition method of creating a matrix with known properties is effected.

G.1.1 Definitions

If A is an $n \times n$ matrix, and i and j are positive integers less than or equal to n , then an $i \times j$ minor (often denoted $M_{i,j}$) of A is the determinant of the $(n-i) \times (n-i)$ matrix obtained from A by deleting from A its i^{th} row and j^{th} column. This means that the minor $M_{3,2}$ of a 3×3 matrix A may be alternatively represented as $det_{r_{1,2}c_{1,3}}$, that is, $M_{3,2} = det_{r_{1,2}c_{1,3}}$.

Let L_d be the lower triangular matrix component of the $L_d.D_d.U_d$ factorisation of A . L_d has only zero-values above its diagonal; all the entries in the first column of L_d are the same as entries in the first column of A and all other non-zero entries are either determinants of submatrices formed with $A_{1,1}$ or minors of A . Each diagonal entry of L_d at position (i) is the determinant of the upper left i -by- i submatrix of A represented as det_i ¹, where $i = 1, 2, \dots, \text{length}(A)$;

$$L_d = \begin{bmatrix} a_{1,1} & 0 & 0 \\ a_{2,1} & det_2 & 0 \\ a_{3,1} & det_{r_{1,3}c_{1,2}} & det_3 \end{bmatrix}, \quad (G.1)$$

Let D_d be a diagonal matrix that is the reciprocal of the product of the determinant of the submatrix formed with $A_{1,1}$ at the i^{th} -position and that of submatrix formed with $A_{1,1}$ from the $(i - 1)^{th}$ point above the diagonal if it exists such that

$$D_d = \begin{bmatrix} \frac{1}{a_{1,1}} & 0 & 0 \\ 0 & \frac{1}{(a_{1,1} * det_2)} & 0 \\ 0 & 0 & \frac{1}{(det_2 * det_3)} \end{bmatrix} \quad (G.2)$$

Let U_d be an upper triangular matrix with all entries below the diagonal set to zero, all entries in the first row of U_d are equivalent to entries in first row of A ; all other entries in U_d are minors of the sub-blocks formed with $A_{1,1}$, such that U_d is of the form:

$$U_d = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ 0 & det_2 & det_{r_{1,2}c_{1,3}} \\ 0 & 0 & det_3 \end{bmatrix} \quad (G.3)$$

We wish to show that

$$A_3 = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix} = L_d * D_d * U_d \quad (G.4)$$

¹ det_{*label} parameter is a determinant or minor of A .

, and that a generalised method for constructing matrices with fixed determinant exists.

G.1.2 Representation of matrix entries by minors

We claim that entries of matrices may be represented by algebraic sums of products of their minors. For example, let A_3 be a 3 x 3 square matrix (with 3x3=9 parameters) such that

$$A_3 = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & \frac{(det_2 + a_{1,2} * a_{2,1})}{a_{1,1}} & \frac{(det_{r_{1,2}c_{1,3}} + a_{1,3} * a_{2,1})}{a_{1,1}} \\ a_{3,1} & \frac{(det_{r_{1,3}c_{1,2}} + a_{1,2} * a_{3,1})}{a_{1,1}} & \frac{r_{3,3}}{(a_{1,1} * det_2)} \end{bmatrix} \quad (G.5)$$

where

$$r_{3,3} = (a_{1,1} * det_3 + det_{r_{1,2}c_{1,3}} * det_{r_{1,3}c_{1,2}} + a_{1,3} * a_{3,1} * det_2).$$

In the same way, A_4 , a 4 x 4 square matrix (with 16 parameters), may be derived. In generalising the algorithm to create square matrices of sizes ≥ 4 , only a slight modification is required. We illustrate how to obtain this generalisation by creating A_4 as an example.

G.1.3 Construction of 4x4 matrices

It is important to show how the $L_d.D_d.U_d$ method may be applied also to the composition or decomposition of a 4x4 matrix. A_4 is derived as a product of $L_d * D_d * U_d$ factors:

$$A_4 = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} \end{bmatrix} = L_d * D_d * U_d$$

$$\text{where } L_d = \begin{bmatrix} a_{1,1} & 0 & 0 & 0 \\ a_{2,1} & \det_2 & 0 & 0 \\ a_{3,1} & \det_{r_{1,3}c_{1,2}} & \det_3 & 0 \\ a_{4,1} & \det_{r_{1,4}c_{1,2}} & \det_{r_{1,2,4}c_{1,2,3}} & \det_4 \end{bmatrix}, D_d = \begin{bmatrix} \frac{1}{a_{1,1}} & 0 & 0 & 0 \\ 0 & \frac{1}{(a_{1,1} * \det_2)} & 0 & 0 \\ 0 & 0 & \frac{1}{(\det_2 * \det_2)} & 0 \\ 0 & 0 & 0 & \frac{1}{(\det_3 * \det_4)} \end{bmatrix}$$

$$\text{and } U_d = \begin{bmatrix} a_{1,1} & a_{2,1} & a_{3,1} & a_{4,1} \\ 0 & \det_2 & \det_{r_{1,2}c_{1,3}} & \det_{r_{1,2}c_{1,4}} \\ 0 & 0 & \det_3 & \det_{r_{1,2,3}c_{1,2,4}} \\ 0 & 0 & 0 & \det_4 \end{bmatrix}.$$

It is not surprising seeing that exactly $4 \times 4 = 16$ distinct parameters are required to construct a 4×4 matrix with a predefined determinant we choose.

G.2 Matrix construction with fixed determinant(s)

In the simulation (artificial creation) of inverse problems for the development and optimisation of inference methods, it is important to ensure that the target jacobian matrix to be inferred is well-conditioned, i.e. it is nonsingular. Nonsingularity may be eliminated by ensuring that the determinant of the matrix is not zero (and not close to it). One of the primary reasons for using nonsingular matrix models in simulating time series data is to ensure that the original jacobian matrix being used is reproducible and can be inferred. Only nonsingular matrices may guarantee that the uniqueness of a potential solution during the inference process is not lost.

G.2.1 Example #1: matrix composition

To construct a matrix with determinant equal to -8, first establish the product of L_d ,

D_d and U_d :

$$L_d = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 5 & -4 & 0 & 0 \\ 9 & -8 & -4 & 0 \\ 13 & -12 & 0 & -8 \end{bmatrix}, D_d = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -\frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{16} & 0 \\ 0 & 0 & 0 & \frac{1}{32} \end{bmatrix}, U_d = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & -4 & -8 & -12 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & -8 \end{bmatrix}$$

The matrix composition method $A_{4a} = L_d \cdot D_d \cdot U_d$ gives the result:

$$A_{4_a} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 12 & 12 \\ 13 & 14 & 15 & 18 \end{bmatrix}$$

G.2.2 Example #2: matrix composition with fewer parameters

To create a matrix with determinant of -8 from fewer nonzero parameters, the requirement is that the parameter det_4 must be set to -8 and that none of the diagonal entries of D_d is zero.

$$A_{4_b} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 5 & -1 & 0 & 0 \\ 6 & 0 & \frac{1}{2} & 0 \\ 7 & 0 & 0 & -8 \end{bmatrix} * \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -\frac{1}{4} \end{bmatrix} * \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & -8 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 9 & 15 & 20 \\ 6 & 12 & \frac{35}{2} & 24 \\ 7 & 14 & 21 & 12 \end{bmatrix}$$

G.2.3 Example #3: symmetric matrix composition

A nonsingular symmetric 4x4 matrix may be constructed by replacing U_d with the transpose of L_d .

$$A_{4s1} = L_d * D_d * L_d^T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 5 & -4 & 0 & 0 \\ 9 & -8 & -4 & 0 \\ 13 & -12 & 0 & -8 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -\frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{16} & 0 \\ 0 & 0 & 0 & \frac{1}{32} \end{bmatrix} \begin{bmatrix} 1 & 5 & 9 & 13 \\ 0 & -4 & -8 & -12 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & -8 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 5 & 9 & 13 \\ 5 & 21 & 37 & 53 \\ 9 & 37 & 66 & 93 \\ 13 & 53 & 93 & 135 \end{bmatrix} \quad (G.6)$$

G.2.4 Example #4: variant symmetric matrix composition

Alternatively, a nonsingular symmetric 4x4 matrix may be constructed by replacing L_d with the transpose of U_d .

$$A_{4s_2} = U_d^T * D_d * U_d \quad (\text{G.7})$$

$$A_{4s_2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & -4 & 0 & 0 \\ 3 & -8 & -4 & 0 \\ 4 & -12 & 0 & -8 \end{bmatrix} * \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -\frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{16} & 0 \\ 0 & 0 & 0 & \frac{1}{32} \end{bmatrix} * \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & -4 & -8 & -12 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & -8 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 0 & -2 & -4 \\ 3 & -2 & -6 & -12 \\ 4 & -4 & -12 & -18 \end{bmatrix}$$

G.2.5 Multiple matrices with a predefined (fixed) determinant

We have demonstrated (using 4 examples) how multiple 4x4 matrices with a fixed determinant of -8 may be constructed using our $L_d.D_d.U_d$ method and showed in the previous section how the following matrices all with the same determinant could be easily created. Specifically A_{4a} , A_{4b} , A_{4s_1} , and A_{4s_2} all have equal determinants, i.e. $\det(A_{4a}) = \det(A_{4b}) = \det(A_{4s_1}) = \det(A_{4s_2}) = -8$. The method, shown here for 4x4 matrices, operates independent of matrix size.

G.3 Decomposition of matrices

G.3.1 $L_d.D_d.U_d$ Decomposition of a Symmetric Matrix

If a square matrix is equal to its transpose then such matrix is said to be symmetric.

If M_s is symmetric, then $M_s = M_s^T$ and is regarded to be of the form

$$M_s = \begin{bmatrix} m_{1,1} & m_{1,2} & m_{1,3} \\ m_{1,2} & m_{2,2} & m_{2,3} \\ m_{1,3} & m_{2,3} & m_{3,3} \end{bmatrix} \quad (\text{G.8})$$

As shown earlier, the $L_d.D_d.U_d$ decomposition of symmetric matrices, e.g. M_s , has the following properties: $L_d^T = U_d$ and $L_d = U_d^T$. This means that the $L_d.D_d.U_d$ decomposition of

$$M_s = L_d.D_d.U_d = U_d^T.D_d.U_d = L_d.D_d.L_d^T \quad (\text{G.9})$$

G.3.2 Relation between Cholesky and $L_d.D_d.U_d$ decomposition methods

The Cholesky decomposition of a symmetric, positive-definite matrix M_s is a factorisation of M_s into $L_c.L_c^*$ where L_c is a lower triangular matrix with positive diagonal entries, and L_c^* is the conjugate transpose of L_c .

From the last definition $M_s = L_d.D_d.U_d = U_d^T.D_d.U_d = L_d.D_d.L_d^T$ we can derive the Cholesky decomposition factors as follows:

$$M_s = L_d * D_d^{1/2} * D_d^{1/2} * L_d^T$$

Since D_d is a diagonal matrix, therefore

$$M_s = L_d * D_d^{1/2} * (D_d^{1/2})^T * L_d^T$$

$$M_s = (L_d * D_d^{1/2}) * (L_d * D_d^{1/2})^T = L_c.L_c^*$$

$$M_s = (U_d * D_d^{1/2})^T * (U_d * D_d^{1/2}) = L_c.L_c^*$$

where $L_c = (L_d * D_d^{1/2}) = (U_d * D_d^{1/2})^T$ for any square matrix M_s . This establishes the relation between our $L_d.D_d.U_d$ decomposition and the Cholesky's.

G.3.3 Relation between LU and $L_d.D_d.U_d$ factorisation methods

Our analysis shows that by solving the system $Ax=b$ with Gaussian elimination, a lower triangular matrix of the form

$$L = \begin{bmatrix} 1 & 0 & 0 \\ \frac{a_{1,2}}{m_{1,1}} & 1 & 0 \\ \frac{a_{1,3}}{m_{1,1}} & \frac{\det_{r_{1,3} c_{1,2}}}{\det_2} & 1 \end{bmatrix}$$

results, which has the same outcome as

$$L = \begin{bmatrix} a_{1,1} & 0 & 0 \\ a_{1,2} & \det_2 & 0 \\ a_{1,3} & \det_{r_{1,3} c_{1,2}} & \det_3 \end{bmatrix} \begin{bmatrix} \frac{1}{a_{1,1}} & 0 & 0 \\ 0 & \frac{1}{\det_2} & 0 \\ 0 & 0 & \frac{1}{\det_3} \end{bmatrix} = L_d * \begin{bmatrix} \frac{1}{m_{1,1}} & 0 & 0 \\ 0 & \frac{1}{\det_2} & 0 \\ 0 & 0 & \frac{1}{\det_3} \end{bmatrix}$$

using our convention. Since

$$L_d = \begin{bmatrix} a_{1,1} & 0 & 0 \\ a_{1,2} & \det_2 & 0 \\ a_{1,3} & \det_{r_{1,3} c_{1,2}} & \det_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{a_{1,2}}{a_{1,1}} & 1 & 0 \\ \frac{a_{1,3}}{a_{1,1}} & \frac{\det_{r_{1,3} c_{1,2}}}{\det_2} & 1 \end{bmatrix} \begin{bmatrix} a_{1,1} & 0 & 0 \\ 0 & \det_2 & 0 \\ 0 & 0 & \det_3 \end{bmatrix} = L * D_m$$

where $D_m = \begin{bmatrix} a_{1,1} & 0 & 0 \\ 0 & \det_2 & 0 \\ 0 & 0 & \det_3 \end{bmatrix}$ is the diagonal matrix containing only the principal minors of the target matrix. Therefore we establish the relation between LU factorisation and our $L_d.D_d.U_d$ decomposition method through the following derivation:

Let $L * D_m = L_d$ where D_m is the right matrix multiplier that transforms the L factor of the well-known LU decomposition to our L_d factor.

Because both the L_d and L factors are derived from the same Gaussian elimination process, it turns out that D_m is a diagonal matrix. Therefore, $L = L_d.D_m^{-1}$.

Since $D_m^{-1}.D_m = I$ where D_m is any square matrix, D_m^{-1} its implied inverse and I

is the identity.

$$L_d.D_d.U_d = L_d.I.D_d.U_d$$

$$L_d.D_d.U_d = L_d.(D_m^{-1}.D_m).D_d.U_d$$

$$L_d.D_d.U_d = (L_d.D_m^{-1}).(D_m.D_d.U_d)$$

Since $L = (L_d.D_m^{-1})$, therefore

$$L_d.D_d.U_d = (L).(D_m.D_d.U_d)$$

And since we want to preserve the integrity of the matrix A in both our decomposition and the LU decomposition processes, the following property holds:

$$A = L_d * D_d * U_d = LU$$

Therefore, LU factorisation itself may be seen to have a new interpretation in terms of our decomposition method, i.e.

$$LU = (L_d.D_m^{-1}) * (D_m.D_d.U_d)$$

$$L = (L_d.D_m^{-1})$$

$$U = (D_m.D_d.U_d)$$

where the matrix D_m is derived to be

$$D_m = \begin{bmatrix} a_{1,1} & 0 & 0 \\ 0 & det_2 & 0 \\ 0 & 0 & det_3 \end{bmatrix}$$

$$L * U = \begin{bmatrix} 1 & 0 & 0 \\ \frac{a_{1,2}}{m_{1,1}} & 1 & 0 \\ \frac{a_{1,3}}{m_{1,1}} & \frac{det_{r_{1,3}c_{1,2}}}{det_2} & 1 \end{bmatrix} * \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ 0 & \frac{det_2}{a_{1,1}} & \frac{det_{r_{1,2}c_{1,3}}}{a_{1,1}} \\ 0 & 0 & \frac{det_3}{det_2} \end{bmatrix} \quad (G.10)$$

where L is lower triangular and U is upper triangular as expected.

G.3.4 Other variants of our $L_d.D_d.U_d$ decomposition method

Other variants of $L_d.D_d.U_d$ decomposition factors may exist in terms of

$$A = L_v.D_v.U_v = \begin{bmatrix} a_{1,1} & 0 & 0 \\ a_{2,1} & 1 & 0 \\ a_{3,1} & \frac{\det_{r_{1,3}c_{1,2}}}{\det_2} & 1 \end{bmatrix} * \begin{bmatrix} \frac{1}{a_{1,1}} & 0 & 0 \\ 0 & \frac{\det_2}{a_{1,1}} & 0 \\ 0 & 0 & \frac{\det_3}{\det_2} \end{bmatrix} * \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ 0 & 1 & \frac{\det_{r_{1,2}c_{1,3}}}{\det_2} \\ 0 & 0 & 1 \end{bmatrix}$$

which has a symmetric matrix equivalent as

$$As = U_{vs}^T.D_{vs}.U_{vs} = \begin{bmatrix} a_{1,1} & 0 & 0 \\ a_{1,2} & 1 & 0 \\ a_{1,3} & \frac{\det_{r_{1,2}c_{1,3}}}{\det_2} & 1 \end{bmatrix} * \begin{bmatrix} \frac{1}{a_{1,1}} & 0 & 0 \\ 0 & \frac{\det_2}{a_{1,1}} & 0 \\ 0 & 0 & \frac{\det_3}{\det_2} \end{bmatrix} * \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ 0 & 1 & \frac{\det_{r_{1,2}c_{1,3}}}{\det_2} \\ 0 & 0 & 1 \end{bmatrix}$$

or

$$As = L_{vs}.D_{vs}.L_{vs}^T = \begin{bmatrix} a_{1,1} & 0 & 0 \\ a_{2,1} & 1 & 0 \\ a_{3,1} & \frac{\det_{r_{1,3}c_{1,2}}}{\det_2} & 1 \end{bmatrix} * \begin{bmatrix} \frac{1}{a_{1,1}} & 0 & 0 \\ 0 & \frac{\det_2}{a_{1,1}} & 0 \\ 0 & 0 & \frac{\det_3}{\det_2} \end{bmatrix} * \begin{bmatrix} a_{1,1} & a_{2,1} & a_{3,1} \\ 0 & 1 & \frac{\det_{r_{1,3}c_{1,2}}}{\det_2} \\ 0 & 0 & 1 \end{bmatrix}$$

Note, these may not be the most reduced forms of factorisation and may require further optimisation of entries.

G.4 Applications of $L_d.D_d.U_d$ method to systems of linear equations

G.4.1 Solving systems of linear systems

The $L_d.D_d.U_d$ decomposition can be applied to solve a system of linear equations such as $A.x = b$ by first computing the $L_d.D_d.U_d$ decomposition of A as $A = (L_d * D_d^{1/2}) * (L_d * D_d^{1/2})^T$ if A is symmetric and positive definite. For example, finding x_u (below) the decomposition can be applied to solving the equation $(L_d * D_d^{1/2}) * x_u = b$ where

$x_u = (L_d * D_d^{1/2})^T * x$. This means that the $L_d.D_d.U_d$ decomposition method may be applied in solving

$$A.x = b$$

problems without having to compute the actual inverse of A.

G.4.2 Application to time series inverse problem analysis

In a system of linear differential equations an inverse problem may be defined in the form

$$\dot{X} = A * X$$

where \dot{X} and X are known vectors of same length, and A is the unknown matrix that must be identified. Note that there is difference between a general system of n linear differential equations with unknown (jacobian) matrix parameters and a general system of n linear equations with unknown vector parameters. The latter is much simpler to solve as explained above. However, the formulation of the inverse problem remains the same in structure. Its algebraic representation is as follows:

$$\begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \\ \vdots \\ \dot{X}_n \end{bmatrix} = \begin{bmatrix} \frac{\partial X_1}{\partial X_1} X_1 + \frac{\partial X_1}{\partial X_2} X_2 \dots + \frac{\partial X_1}{\partial X_m} X_m \\ \frac{\partial X_2}{\partial X_1} X_1 + \frac{\partial X_2}{\partial X_2} X_2 \dots + \frac{\partial X_2}{\partial X_m} X_m \\ \vdots \\ \frac{\partial X_m}{\partial X_1} X_1 + \frac{\partial X_m}{\partial X_2} X_2 \dots + \frac{\partial X_m}{\partial X_m} X_m \end{bmatrix} \quad (G.11)$$

where the partial derivative parameters are the entries of the jacobian matrix A; A contains the relative rates of change with respect to the dependent variables (Gilbert 1988). In light of this definition, a time series inverse problem may be defined (in a mathematical sense) as a general system of m linear differential equations with an unknown m x m jacobian matrix A may be rewritten in the form $\dot{X} = A.X$, which

has the matrix equation form

$$\begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \\ \vdots \\ \dot{X}_m \end{bmatrix} = \begin{bmatrix} \frac{\partial X_1}{\partial X_1} & \frac{\partial X_1}{\partial X_2} & \cdots & \frac{\partial X_1}{\partial X_m} \\ \frac{\partial X_2}{\partial X_1} & \frac{\partial X_2}{\partial X_2} & \cdots & \frac{\partial X_2}{\partial X_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial X_m}{\partial X_1} & \frac{\partial X_m}{\partial X_2} & \cdots & \frac{\partial X_m}{\partial X_m} \end{bmatrix} \cdot \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{bmatrix} \quad (\text{G.12})$$

where $\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{bmatrix} = X(t)$ is a known state vector, i. e. the t^{th} vector of the given time

series X. Rewritten in a multi-state definition, if m number of (state) measurements are taken after the initial condition, it becomes

$$\begin{bmatrix} \dot{X}_{10} \dot{X}_{11} \cdots \dot{X}_{1n-1} \\ \dot{X}_{20} \dot{X}_{21} \cdots \dot{X}_{2n-1} \\ \vdots \\ \dot{X}_{m0} \dot{X}_{m1} \cdots \dot{X}_{mn-1} \end{bmatrix} = \begin{bmatrix} \frac{\partial X_1}{\partial X_1} & \frac{\partial X_1}{\partial X_2} & \cdots & \frac{\partial X_1}{\partial X_m} \\ \frac{\partial X_2}{\partial X_1} & \frac{\partial X_2}{\partial X_2} & \cdots & \frac{\partial X_2}{\partial X_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial X_m}{\partial X_1} & \frac{\partial X_m}{\partial X_2} & \cdots & \frac{\partial X_m}{\partial X_m} \end{bmatrix} * \begin{bmatrix} X_{10} X_{11} \cdots X_{1n-1} \\ X_{20} X_{21} \cdots X_{2n-1} \\ \vdots \\ X_{m0} X_{m1} \cdots X_{mn-1} \end{bmatrix}$$

which is equivalent to

$$\begin{bmatrix} \frac{X_{11}-X_{10}}{t_c} \frac{X_{12}-X_{11}}{t_c} \cdots \frac{X_{1n}-X_{1n-1}}{t_c} \\ \frac{X_{21}-X_{20}}{t_c} \frac{X_{22}-X_{21}}{t_c} \cdots \frac{X_{2n}-X_{2n-1}}{t_c} \\ \vdots \\ \frac{X_{m1}-X_{m0}}{t_c} \frac{X_{m2}-X_{m1}}{t_c} \cdots \frac{X_{mn}-X_{mn-1}}{t_c} \end{bmatrix} = \begin{bmatrix} \frac{\partial X_1}{\partial X_1} & \frac{\partial X_1}{\partial X_2} & \cdots & \frac{\partial X_1}{\partial X_m} \\ \frac{\partial X_2}{\partial X_1} & \frac{\partial X_2}{\partial X_2} & \cdots & \frac{\partial X_2}{\partial X_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial X_m}{\partial X_1} & \frac{\partial X_m}{\partial X_2} & \cdots & \frac{\partial X_m}{\partial X_m} \end{bmatrix} \cdot \begin{bmatrix} X_{10} X_{11} \cdots X_{1n-1} \\ X_{20} X_{21} \cdots X_{2n-1} \\ \vdots \\ X_{m0} X_{m1} \cdots X_{mn-1} \end{bmatrix}$$

assuming that the state vector measurements are captured at regular time intervals

of t_c . Note that $e^{t_c} \approx 1 + t_c$, of $e^{t_c} = 1 + t_c + \frac{t_c^2}{2!} + \frac{t_c^3}{3!} + \dots$ improves as t_c gets closer

to 0 (Gilbert 1988). The solution to this system of linear differential equations is

$$\begin{bmatrix} X_{1_1} X_{1_2} \dots X_{1_n} \\ X_{2_1} X_{2_2} \dots X_{2_n} \\ \vdots \\ X_{m_1} X_{m_2} \dots X_{m_n} \end{bmatrix} = \exp \begin{bmatrix} \frac{\partial X_1}{\partial X_1} & \frac{\partial X_1}{\partial X_2} \dots & \frac{\partial X_1}{\partial X_m} \\ \frac{\partial X_2}{\partial X_1} & \frac{\partial X_2}{\partial X_2} \dots & \frac{\partial X_2}{\partial X_m} \\ \vdots & \vdots & \vdots \\ \frac{\partial X_m}{\partial X_1} & \frac{\partial X_m}{\partial X_2} \dots & \frac{\partial X_m}{\partial X_m} \end{bmatrix} . t_c * \begin{bmatrix} X_{1_0} X_{1_1} \dots X_{1_{n-1}} \\ X_{2_0} X_{2_1} \dots X_{2_{n-1}} \\ \vdots \\ X_{m_0} X_{m_1} \dots X_{m_{n-1}} \end{bmatrix}$$

Further analysis of such inverse problems might require matrix decomposition, e.g., the system above may be redefined as

$$X_{(after)} = \exp^{[A * t_c]} * X_{(before)}$$

G.4.3 Solving time series inverse problem using matrix manipulation

In solving time series inverse problems of the form

$$X_{(after)} = \exp^{[A * t_c]} * X_{(before)}$$

where the jacobian matrix A must be identified and t_c is the time interval of very small magnitude, Idowu and Bown (Idowu M.A. 2011a) developed the transposive regression method for finding $E = \exp^{[A * t_c]}$.

G.5 Conclusion: matrix construction and decomposition method

We have demonstrated that through simple matrix manipulation techniques, it is possible to derive methods for creating and deconstructing matrices with known determinants. A robust matrix-based framework to enable simplified creation and evaluation

of system identification and parameter estimation problems and solutions is being developed, and the method presented here is an essential part of that framework. This framework, an important tool for developing and managing optimisation methods, is required in systems biology and applicable to other areas such as artificial intelligence, network science, etc. The inference method is sophisticated enough for the identification of ODE models from time series data. In future work, we will demonstrate how these matrix decomposition and composition methods presented here may be applied to the development of new techniques for understanding complex network structures.

Bibliography

- Adjei A.A., H. M. (2005), 'Intracellular signal transduction pathway proteins as targets for cancer therapy', *J Clin Oncol* . 23: 5386-5403.
- Akutsu T., Miyano S., K. S. (2000), 'Inferring qualitative relations in genetic networks and metabolic pathways', *Bioinformatics* . Vol. 16 no. 8, 727-734, doi: 10.1093/bioinformatics/16.8.727.
- Alan, W. (1999), 'Egf receptor', *Int J Biochem Cell Biol* . 31:637-643.
- Albert R., B. A.-L. (2001), 'Statistical mechanics of complex networks'.
- Alberts Bruce, Johnson Alexander, L. J.-R. M. R. K. W. P. (2009), *Molecular biology of the cell*, 5th edn.
- Aldridge B.B., Burke J.M., L. D.-S. P. (2006), 'Physiochemical modelling of cell signalling pathways', *Nature Cell Biology* . 8(11):1195-1203.
- Alexander Kamb, S. W. & Lengauer, C. (2007), 'Why is cancer drug discovery so difficult', *Nature Reviews Drug Discovery* . 6, 115-120, doi:10.1038/nrd2155.
- Almeida J.S., V. E. (2003), 'Neural-network-based parameter estimation in s-system models of biological networks', *Genome Informatics* . 14: 114-123.
- Alves R., S. M. (2000), 'Extending the method of mathematically controlled comparison to include numerical comparisons', *Vol. 16 n0. 9* . Pages 786-798.
- A.M, I. (2007), 'Personalized medicine and the practice of medicine in the 21st century', *Mcgill J Med* . 10(1): 53-57.

- Amit I., Wides R., Y. Y. (2007), 'Evolvable signaling networks of receptor tyrosine kinases:relevance of robustness to malignancy and to cancer therapy', *Molecular Systems Biology* . (3):151.
- Andrea Sackmann, M. H. & Koch, I. (2006), 'Application of petri net based analysis techniques to signal transduction pathways', *BMC Bioinformatics* . 7:482, doi:10.1186/1471-2105-7-482.
- A.R., A. (2008), 'Quaranta v. integrative mathematical oncology', *Nat Rev Cancer* . 8(3):227-34.
- Araujo R.P., Liotta L.A., P. E. (2007), 'Proteins, drug targets and the mechanisms they control: the simple truth about complex networks', *Nature Reviews Drug Discovery* . 6, 871-880, doi:10.1038/nrd2381.
- Arne, B. (1951), 'Application of calculus of matrices to method of least squares; with special references to geodetic calculations', *Trans. Roy. Inst. Tech. Stockholm* 49 .
- A.S, H. (1975), 'The theory of matrices in numerical analysis', *New York Dover Publications* . ISBN 0486617815.
- Axler, S. (1995), 'Down with determinants!', *American Mathematical Monthly* 102 . 139-154.
- B, K. (2006), 'Cell signalling dynamics in time and space', *Nature Reviews Molecular Cell Biology* . 7(3):165-176.
- Bailey W.J., U. R. (2004), 'Molecular profiling approaches for identifying novel biomarkers', *Expert Opin. Drug Saf* . 3(2):137-151.
- Bansal M., Belcastro V., A.-I. A. d.-B. D. (2007), 'How to infer gene networks from expression profiles', *Molecular Systems Biology* . 13 February, doi:10.1038/msb4100120.

- Bansal M, Della Gatta G, d. B. D. (2006), 'Inference of gene regulatory networks and compound mode of action from time course gene expression profiles', *Bioinformatics* 22: 815-822 .
- Bansal M., Della Gatta G., W. J. G. T. d. B. D. (2005), 'Conf. proc. ie eng. med. biol', *Soc.*, 5, 4739-4742 .
- Bown J., Andrews P.S., D. Y. G. A. I. M. P. F. S. A. S. M. S. S. (2012), 'Engineering simulations for cancer systems biology', *Current Drug Targets* .
- Brazhnik P., Fuente A., M. P. (2002), 'Gene networks: how to put the function in genomics, trends biotechnol', 20(11):467-472 .
- Breathrough (2010), 'www.breakthrough.org.uk/about_breast_cancer/diagnosis/', *Breakthrough website* . Last accessed: 2010.
- Briggs G.E., H. J. (1925), 'A note on the kinematics of enzyme action', *Biochem J* 19 (2): 338-339 . PMC 1259181, PMID 16743508.
- Brown K.S., Hill C.C., C. G. M. C. L. K. S. J. C. R. (2004), 'The statistical mechanics of complex signaling networks: nerve growth factor signaling', *Phys Biol* 1:184 .
- B.S, H. (2005), 'Parsing erk activation reveals quantitatively equivalent contributions from epidermal growth factor receptor and her2 in human mammary epithelial cells', *J. Biol. Chem* . 280, 6157-6169.
- BSTLab (2007), 'A preliminary web-based application for ann smoothing is accessible a bioinformatics.musc.edu/webmetabol/. freeware plas with matlab module bstlab', *MUSC Bioinformatics website* . <http://correio.cc.fc.ul.pt/aenf/plas.html>, Last accessed: 2007, <https://bioinformatics.musc.edu/bstlab/>.
- Burke J.V., Lewis A.S., O. M. (2003), 'Robust stability and a criss-cross algorithm for pseudospectra', *IMA Journal of Numerical Analysis* . 23 (3), pp 359-375, doi: 10.1093/imanum/23.3.359.

- Butcher E.C., Berg E.L., K. E. (2004), 'Systems biology in drug discovery', *Nature Biotechnology* 22, 1253 - 1259 (2004) . Published online: 6 October, doi:10.1038/nbt1017.
- C, A. (2003), 'Targeting her1/egfr: a molecular approach to cancer therapy', *Semin Oncol* . 30:3-14.
- Carmen G. Moles, P. M. & Banga, J. R. (2003), 'Parameter estimation in biochemical pathways: A comparison of global optimization methods', 13:2467-2474 . www.genome.org/cgi/doi/10.1101/gr.1262503.
- CellLectures (2010), 'www.fsm.ac.fj/pws/resources/lectures/', *FSM website* . Last accessed: 2010.
- Chen L., Wang R.S., Z. X. (2009), 'Biomolecular networks. methods and applications in systems biology', *John Wiley and Sons* .
- Chen W.W., Schoeberl B., J. P. N. M. N. U. (2009), 'Input-output behaviour of erbb signaling pathways as revealed by a mass action model trained against dynamic data', *Molecular Systems Biology* . (5):239.
- Chih-Lung Ko, Voit E.O., W. F.-S. (2009), 'Estimating parameters for generalized mass action models with connectivity information', *BMC Bioinformatics* . 10:140, Last accessed: 2007, <https://bioinformatics.musc.edu/bstlab/>, doi:10.1186/1471-2105-10-140.
- Chou I-Chun, Martens H., V. E. (2006), 'Parameter estimation in biochemical systems models with alternating regression', *Theoretical Biology and Medical Modelling* 2006 . 3:25, doi:10.1186/1742-4682-3-25.
- Chou I-Chun, Martens H., V. E. (2007), 'Parameter estimation of s-distributions with alternating regression', *SORT* 31 (1) . 55-74.
- Chou I-Chun, V. E. (2009), 'Recent developments in parameter estimation and structure identification of biochemical and genomic systems', *Mathematical Biosciences* 219 . 57-83, doi:10.1016/j.mbs.20.

- Chua H.N., R. F. (2011), 'Discovering the targets of drugs via computational systems biology', *J. Biol. Chem.* . 286, 23653-23658.
- Citri A., Y. Y. (2006), 'Egf-erbB signalling: towards the systems level', *Nature Reviews Molecular Cell Biology* . (7):505-516.
- Date, D. D. R. (2003), 'First printed in r and d systems', *Catalog* .
- David Gilbert, Hendrik Fu, X. G. (2006), 'Computational methodologies for modelling, analysis and simulation of signalling networks', *Brief Bioinform.* 7(4):339-353.
- Dawson J.P., Berger M.B., L. C. S. J. L. M. F. K. (2005), 'Epidermal growth factor receptor dimerization and activation require ligand-induced conformational changes in the dimer interface', *Mol Cell Biol.* 25:7734-7742.
- Derek Ruths, Luay Nakhleh, P. T. R. (2008), 'Rapidly exploring structural and dynamic properties of signaling networks using pathwayoracle', *BMC Systems Biology* . 2:76, doi:10.1186/1752-0509-2-76.
- D.H, I. (1988), 'Efficient solution of nonlinear models expressed in s-system canonical form', *Math1 Comput. Modelling* . Vol. I I, pp. 123-128.
- Diaz-Sierra R., F. V. (2001), 'Simplified method for the computation of parameters of power-law rate equations from time series', *Mathematical Biosciences* . 171, 1-19.
- DiMasi J.A., Hansen R.W., G. H. (2003), 'The price of innovation: New estimates of drug development costs', *Journal of Health Economics* pp. 835,1-35.
- D.M, B. (2006), 'Causes of cancer', *Professor of Biology* . Saint Michael's College.
- Dorsey E.R., Thompson J.P., C. M. d. R. J. V. P. e. a. (2009), 'Financing of u.s. biomedical research and new drug approvals across therapeutic areas', *PLoS ONE* . 4(9): e7015, doi:10.1371/journal.pone.0007015.

- drug shortages hit vast majority of oncologists, C. (2013), ‘Oncologynurseadvisor’, *www.oncologynurseadvisor.com/cancer-drug-shortages-hit-vast-majority-of-oncologists/article/298527/#* . Last accessed: 19/06/2013.
- E, G. M. M. (2004), ‘Modelling the activity of single genes, in computational modelling of genetic and biochemical networks’, *Bower, J. M., bolou, H. (eds.)* . MIT press, Cambridge, MA.
- Ebenbauer, C. (2007), ‘A dynamical system that computes eigenvalues and diagonalizes matrices with a real spectrum’, *In Proc. of the 46th IEEE Conference on Decision and Control (CDC)* . New Orleans, USA, pages 1704-1709.
- E.H, M. (1920), ‘On the reciprocal of the general algebraic matrix’, *Bulletin of the American Mathematical Society* 26: 394-395 . projecteuclid.org/euclid.bams/1183425340.
- E.K, R. (2003), ‘Signal events: cell signal transduction and its inhibition in cancer’, *Oncologist* . 8(suppl 3):5-17.
- Faratian D., Goltsov A., L. G. M. S. M. P. K. C. H. U. I. L. S. G. I. & Harrison, D. J. (2009), ‘Systems biology reveals new strategies for personalising cancer medicine and confirms pten’s role in resistance to trastuzumab’, *Cancer Research*, (69):6713 .
- FDA (2012), ‘Bringing life-saving drugs to patients quickly and efficiently’, *FY 2012 Innovative Drug Approvals* . FDA. U.S. Department of Health and Human Services. U.S. Food and drug administration (www.fda.gov).
- Gardner T.S., Bernardo, D. L. D. C. J. (2003), ‘Inferring genetic networks and identifying compound mode of action via expression profiling’, *Science* 301:102-105 .
- Gennemark P., W. D. (2009), ‘Benchmarks for identification of ordinary differential equations from time series data’, *Bioinformatics* . Vol. 25 no. 6, pages 780-786, doi:10.1093/bioinformatics/btp050.

- Gilbert, S. (1988), 'Linear algebra and its applications', *Third edition* . ISBN: 0155510053.
- Goel G., Chou I-Chun, V. E. (2008), 'System estimation from metabolic time series data', *Bioinformatics Vol. 24 no. 21* . pages 2505-2511, doi:10.1093/bioinformatics/btn470.
- Goltsov A., Faratian D., L. S. B. J. G. I. & Harrison, D. (2011), 'Compensatory effects in the pi3k/pten/akt signaling network following receptor tyrosine kinase inhibition', *Cellular Signalling* . (23):407-416.
- Gutenkunst R.N., Waterfall J.J., C. F. B. K. M. C. e. a. (2007), 'Universally sloppy parameter sensitivities in systems biology models. plos comput biol', *3(10): e189*. doi:10.1371/journal.pcbi.0030189 .
- H, K. (2002a), 'Computational systems biology', *Nature* . 420:206.
- H, K. (2002b), 'Systems biology: A brief overview', *Science* . (295):1662-1664.
- H, K. (2004), 'Cancer as a robust system: implications for anticancer therapy', *Nature Reviews Cancer* 4 . 227-235.
- H, K. (2007), 'A robustness-based approach to systems-oriented drug design', *Nature Reviews Drug Discovery* 6 . 202-210.
- Hanahan D., W. R. (2000), 'The hallmarks of cancer', *Cell* . Volume 100, Issue 1, 7 January, Pages 57-70, ISSN 0092-8674, doi:10.1016/S0092-8674(00)81683-9, www.sciencedirect.com/science/article/pii/S0092867400816839.
- Hernandez-Bermejo B., Fairen V., S. A. (1999), 'Power-law modeling based on least-squares minimization criteria', *Mathematical Biosciences* 161 . 83-94.
- Hernandez-Bermejo B., Fairen V., S. A. (2000), 'Power-law modeling based on least-squares criteria: consequences for system analysis and simulation', *Mathematical Biosciences* 167 . 87-107.

- Ideker T., Galitski T., H. L. (2001), 'A new approach to decoding life: Systems biology', *Annu. Rev. Genomics Hum. Genet.* . 2:343.
- Ideker T., L. D. (2003), 'Building with a scaffold: emerging strategies for high to low-level cellular modeling', *Trends Biotechnol.* 21, 255-262 .
- Ideker T.L., Winslow R., L. D. (2006), 'Bioengineering and systems biology', *Annals of Biomedical Engineering* . Vol. 34, No. 2, pp. 257-264, doi: 10.1007/s10439-00590477.
- Idowu M.A., B. J. (2011a), 'Towards an exact reconstruction of a time-invariant model from time series data', *Journal of Comp. Sci. and Syst. Biol.* . vol. 4, pp. 055-070, doi:10.4172/jcsb.1000077.
- Idowu M.A., B. J. (2012), 'Matrix operations for the simulation and immediate reverse-engineering of time series data', *Proceedings - 2012 14th International Conference on Modelling and Simulation, UKSim 2012* . art. no. 6205435, pp. 101-106.
- Idowu M.A., B. J. (2013), 'Matrix-based analytical methods for recasting jacobian models to power-law models', *Computer Modelling and Simulation (EuroSim), 8th EUROSIM Congress on Modelling and Simulation* . Pages 250 - 258, doi: 10.1109/EUROSIM.2013.53.
- Idowu M.A., Goltsov A., K. H. T. H. Z. N. B. J. (2011b), 'Cancer research and personalised medicine: a new approach to modelling time series data using analytical methods and half systems', *Current Opinion in Biotechnology* . Volume 22, Supplement 1, Page S59,.
- ImpactReport (2002), 'Tufts csdd quantifies savings from boosting new drug r and d efficiency, tufts center for the study of drug development', *IMPACT Report* . Vol 4, No 5, September,October.
- J, A. (2003), '<http://www.nigms.nih.gov/funding/systems.html>'.

- Janes K.A., Y. M. (2006), 'Data-driven modelling of signal-transduction networks', *Nature Reviews Molecular Cell Biology* 7, 820-828 . doi:10.1038/nrm2041.
- J.K, N. (2006), 'Global systems biology, personalized medicine and molecular epidemiology', *Molecular Systems Biology* 2 Article number: 52 . doi:10.1038/msb4100095.
- Jones J.T., Akita R.W., S. M. (1999), 'Binding specificities and affinities of egf domains for erbb receptors', *FEBS Lett* . 447:227-231.
- J.S, A. (2002), 'Predictive non-linear modeling of complex data by artificial neural networks', *Current Opinion in Biotechnology* . 13:72-76.
- Kholodenko B.N., Hancock J.F., K. W. (2010), 'Signalling ballet in space and time', *Nature Reviews Molecular Cell Biology* . (11):414-426.
- Kikuchi S., Tominaga D., A. M. T. K. T. M. (2003), 'Dynamic modeling of genetic networks using genetic algorithm and s-system', *Vol. 19 no. 5* . pages 643-650, doi:10.1093/bioinformatics/btg027.
- Kim H.H., Sierke S.L., K. J. (1994), 'Epidermal growth factor-dependent association of phosphatidylinositol 3-kinase with the erbb3 gene product', *J. Biol. Chem.* 269 . 24747-24755.
- Kitayama T., Kinoshita A., S. M. N. Y. T. M. (2006), 'A simplified method for power-law modelling of metabolic pathways from time-course data and steady-state flux profiles', *Theoretical Biology and Medical Modelling* . 3:24, doi:10.1186/1742-4682-3-24.
- Kreeger P.K., L. D. (2009), 'Cancer systems biology: a network modeling perspective', *Carcinogenesis* . 31(1):2-8.
- Kreeger P.K., L. D. (2010), 'Cancer systems biology: a network modeling perspective carcinogenesis', *31(1): 2-8 first published online October 27, 2009* . doi:10.1093/carcin/bgp261.

- Krzysztof Fujarewicz, Marek Kimmel, T. L. & Swierniak, A. (2007), 'Adjoint systems for models of cell signaling pathways and their application to parameter fitting', *IEEE/ACM Transactions on Computational Biology and Bioinformatics* . Vol. 4, No. 3.
- Liao J.C., Boscolo R., Y. Y. T. L. S. C. e. a. (2003), 'Network component analysis: reconstruction of regulatory signals in biological systems', *PNAS* 100:15522-15527 .
- M.B, K. (2008), 'Dna damage responses: Mechanisms and roles in human disease', *Molecular cancer research* 6.
- McCubreya J.A., S. L. (2006), 'Advan. enzyme regul', 46 249 .
- Morris M.K., Saez-Rodriguez J., S. P. L. D. (2010), 'Logic-based models for the analysis of cell signaling networks', *Biochemistry* 49(15), 3216-3224 . PMID: 20225868, doi:10.1021/bi902202q, www.ncbi.nlm.nih.gov/pubmed/20225868.
- Naidu R., Yadav M., N. S. K. M. (1998), 'Expression of c-erbB3 protein in primary breast carcinomas', *Br J Cancer* 78: 1385-1390B .
- Nasimul Noman, H. I. (2005), 'Inference of gene regulatory networks using s-system and differential evolution', *Proceeding GECCO '05 Proceedings of the 2005 conference on Genetic and evolutionary computation* . ACM New York, NY, USA, ISBN1-59593-010-8, doi10.1145/1068009.1068079.
- Nasimul Noman, H. I. (2006), 'Inference of genetic networks using s-system: Information criteria for model selection', *GECCO '06 Proceedings of the 8th annual conference on Genetic and evolutionary computation* . ACM New York, NY, USA. ISBN:1-59593-186-4, doi:10.1145/1143997.1144043.
- Nemenman I., Escola G.S., H. W. U. P. U. C. W. M. (2007), 'Reconstruction of metabolic networks from high-throughput metabolite profiling data', *In Silico Analysis of Red Blood Cell Metabolism* . Ann. N.Y. Acad. Sci. 1115: 102-115, doi: 10.1196/annals.1407.013.

- Ni Ta-chen, S. M. (1996), 'Model assessment and refinement using strategies from biochemical systems theory', *Application to Metabolism in Human Red Blood Cells* . J. theor. Biol, 179, 329-368.
- Nyarko E.K., S. R. (2004), 'Solving the parameter identification problem of mathematical models using genetic algorithms', *Applied Mathematics and Computation* 153 . 651-658.
- Oda K., Matsuoka Y., F. A. K. H. (2005), 'A comprehensive pathway map of epidermal growth factor receptor signaling', *Mol Syst* . Biol 1: 2005.0010.
- Olayioye M.A., Neve R.M., L. H. H. N. (2000), 'The erbb signalling network: receptor heterodimerization in development and cancer', *Embo J* . 19:3159-3167.
- Orland R. Gonzalez, Christoph Kuper, K. J. P. C. N. J. E. M. (2006), 'Parameter estimation using simulated annealing for s-system models of biochemical networks', *Bioinformatics* (2007) . 23 (4): 480-486, doi: 10.1093/bioinformatics/btl522.
- Papin J.A., Hunter T., P. B. S. S. (2005), 'Reconstruction of cellular signalling networks and analysis of their properties', *Nature Reviews Molecular Systems Biology* . (6):99-111.
- Polisetty P.K., Voit E.O., G. E. (2006), 'Identification of metabolic system parameters using global optimization methods', *Theoretical Biology and Medical Modelling* 2006 . 3:4, doi:10.1186/1742-4682-3-4.
- Roger, P. (1955), 'A generalized inverse for matrices', *Proceedings of the Cambridge Philosophical Society* . 51: 406-413, doi:10.1017/S0305004100030401.
- Rosario R.C.H.d., Mendoza E., V. E. (2008), 'Challenges in lin-log modelling of glycolysis in lactococcus lactis', *IET Syst. Biol* . Pages 136-149, Vol.2 No. 3, ISSN:1751-8849 [ID No:2], Doi 367816, 2008.
- Salvatore Pece, Daniela Tosoni, S. C. G. M. M. V. S. R. L. B. G. V. P. G. P. & Fiore, P. P. D. (2010), 'Biological and molecular heterogeneity of breast cancers correlates with their cancer stem cell', *content. Cell*, 140(1):62-73 .

- Savageau, A. M. (1969), 'Biochemical systems analysis', *J. Theor. Biol.* 25 . 365-379.
- Schoeberl B., Pace E.A., F. J. H. B. X. L. N. L. L. B. K. A. P. V. B. R. G. V. K. N. W. K. L. M. F. M. K. A. N. U. (2009), 'Therapeutically targeting erbb3: A key node in ligand-induced activation of the erbb receptor-pi3k axis', *Science Signalling* . 2, 2(77):ra31.
- Searson D.P., Willis M.J., H. S. W. A. (2007), 'S-systems and evolutionary algorithms for the inference of chemical reaction networks from fed-batch reactor experiments', *Chemical Product and Process Modeling* . Volume 2, Issue 1, Article 10.
- Shovman M., Idowu M., G. A. B. J. (2010), 'Dynamic visualisation of biological network models', *Intl. Conf. on Systems Biology, Edinburgh* .
- Shuhei Kimura, Kaori Ide, A. K. M. K. M. H. R. M. N. N. S. Y. S. K. & Konagaya, A. (2005), 'Inference of s-system models of genetic networks using a cooperative coevolutionary algorithm', *Vol. 21 no. 7* . pages 1154-1163,.
- Slamon D.J., Clark G.M., W. S. L. W. U. A. M. W. (1987), 'Human breast cancer: correlation of relapse and survival with amplification of the her-2/neu oncogene', *Science* 235: 177-182 .
- Slamon D.J., Godolphin W., J. L. H. J. W. S. K. D. L. W. S. S. U. J. U. A. (1989), 'Studies of the her-2/neu protooncogene in human breast and ovarian cancer', *Science* 244: 707-712 .
- Software, G. G. V. (2011), 'www.graphviz.org', *Graphviz.org* . Last accessed: 2013.
- Soltoff S.P., Carraway K.L.III, P. S. G. W. C. L. (1994), 'ErbB3 is involved in activation of phosphatidylinositol 3-kinase by epidermal growth factor', *Mol. Cell. Biol.* 14, 3550-3558.
- Sorribas A., C. M. (1994), 'Structure identifiability in metabolic pathways: parameter estimation in models based on the power-law formalism', *Biochem. J.* 298, 303-311 .

- Spieth C., Streichert F., S. N. Z. A. (2004), 'A memetic inference method for gene regulatory networks based on s-systems', *In Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2004)* . pages 152-157, CiteSeerX 10.1.1.70.7558.
- Srividhya J., Crampin E.J., M. P. S. S. (2007), 'Reconstructing biochemical pathways from time course data', *Proteomics*, 7 . 828-838, doi:10.1002/pmic.200600428.
- Stephens P., Hunter C., B. G. E. S. D. H. e. e. (2004), 'Lung cancer: intragenic erbb2 kinase mutations in tumors', *Nature* 431:525-526 .
- Tominaga D., O. M. (1998), 'Design of canonical model describing complex nonlinear dynamics', *Proc. IFAC Int. Conf.* . CAB785-90.
- Torres N.V., Alvarez-asquez F., V. E. (2003), 'Introduction to the theory of metabolic modeling and optimization of biochemical systems', *Application to Citric Acid Production in Aspergillus niger* . 17/3/2003-SHARON-027.
- Tournier, L. (2005), 'Approximation of dynamical systems using ssys-tems theory: Application to biological systems', *ISBN:1-59593-095-7* . doi:10.1145/1073884.1073928.
- Tsai K.Y., W. F. (2005), 'Evolutionary optimization with data collocation for reverse engineering of biological networks', *Bioinformatics* . vol. 21, no. 7, pages 1180-1188, doi:10.1093/bioinformatics/bti099.
- UK, C. (2012), 'Cancerstats report - cancer in the uk: 2011 december', *Cancer Research UK* . info.cancerresearchuk.org/cancerstats/index.htm, Last assessed: May 22.
- Veflingstad S.R., Almeida J., V. E. (2004), 'Priming nonlinear searches for pathway identification', *Theoretical Biology and Medical Modelling* . 1:8, doi:10.1186/1742-4682-1-8.

- Vilela M., Chou I-Chun, V. S. T. A. V. R. V. E. A. J. (2008), 'Parameter optimization in s-system models', *BMC Systems Biology* . 2:35 doi:10.1186/1752-0509-2-35, www.biomedcentral.com/1752-0509/2/35.
- Vilela M., Vinga S., G. M. M. M. V. E. A. J. (2009), 'Identification of neutral biochemical network models from time series data', *BMC Systems Biology* . 3:47, doi:10.1186/1752-0509-3-47, www.biomedcentral.com/1752-0509/3/47.
- Voit, E. O. (1991), 'Canonical nonlinear modelling', *S-system Approach to Understanding Complexity* . (ed.), xi + 365 pp, Van Nostrand Reinhold.
- Voit, E. O. (2000), 'Computational analysis of biochemical systems', *A practical guide for biochemists and molecular biologists* . xii + 530 pp, Cambridge University Press, Cambridge, U.K.
- Voit, E. O. (2002a), 'Metabolic modeling: a tool of drug discovery in the post-genomic era', *Research focus reviews* . DDT Vol. 7, No. 11.
- Voit, E. O. (2002b), 'Models-of-data and models-of-processes in the post-genomic era', *Mathematical Biosciences* 180 . 263-274.
- Voit, E. O. (2004), 'Biochemical systems theory', *The Integration of Chemical and Biological Engineering Workshop B: Kinetics and Reactor Engineering* .
- Voit, E. O. (2005), 'Smooth bistable s-systems', *Syst Biol (Stevenage)* . 152(4):207-13.
- Voit, E. O. (2008), 'Model identification: A key challenge is computational systems biology', *The Second International Symposium on Optimization and Systems Biology (OSB-08)* . Lijiang, China, ORSC and APORC, pp. 1-12.
- Voit, E. O. (2009), 'Parameter estimation and structure identification in metabolic pathway systems', *Workshop on Parameter Estimation for Dynamical Systems* . EURANDOM, Eindhoven, The Netherlands, June 8-10.
- Voit, E. O. (2013), 'Biochemical systems theory: A review', *ISRN Biomathematics* . vol. 2013, Article ID 897658, 53 pages, doi:10.1155/2013/897658.

- Voit E.O., A. J. (2004a), 'Decoupling dynamical systems for pathway identification from metabolic profiles bioinformatics', *Jul 22;20(11):1670-81* . <https://bioinformatics.musc.edu/webmetabol/>, free-ware PLAS, <http://correio.cc.fc.ul.pt/aenf/plas.html>, module BSTLab <https://bioinformatics.musc.edu/bstlab/>.
- Voit E.O., C. I.-C. (2010), 'Parameter estimation in canonical biological systems models', *International Journal of Systems and Synthetic Biology* . pp. 1-19.
- Voit E.O., Goel G., C. I.-C. F. L. (2009), 'Estimation of metabolic pathway systems from different data sources', *IET Systems Biology* . Special Issue - Selected papers from The 2nd International Symposium on Optimization and Systems Biology (OSB 2008), ISSN 1751-8849, doi: 10.1049/iet-syb.2008.0180.
- Voit E.O., Marino S., L. R. (2004b), 'Challenges for the identification of biological systems from in vivo time series data', *In Silico Biology 5, 0010* . Bioinformation Systems e.V.
- Voit E.O., R. T. (2000), 'Biochemical systems analysis of genome-wide expression data', *vol. 16 no. 11* . Bioinformatics, 16 (11): 1023-1037, doi: 10.1093/bioinformatics/16.11.1023.
- Warwick Tucker, V. M. (2006), 'Parameter reconstruction for biochemical networks using interval analysis reliable computing', *12: 389-402* . doi:10.1007/s11155-006-9009-2.
- W.J, G. (2001), 'The type 1 growth factor receptors and their ligands considered as a complex system', *Endocr Relat Cancer* . 8:75-82.