CD-ROM ON BACK COVER

**NOTICE OF SUBMISSION OF THESIS FORM:
POSTGRADUATE RESEARCH**

CARDIFF
UNIVERSITY
PRIFYSGOL
CAERDYDD

APPENDIX 1:
Specimen layout for Thesis Summary and Declaration/Statements page to be included in a Thesis

DECLARATION

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed .... (candidate) Date 24/06/08..............

STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of ....PhD..................(insert MCh, MD, MPhil, PhD etc, as appropriate)

Signed .... (candidate) Date 24/06/08..............

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed .... (candidate) Date 24/06/08..............

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .... (candidate) Date 24/06/08..............

A novel bioinformatics approach for encoding and interrogating the progression and modulation of the mammalian cell cycle

Imtiaz Ali Khan

Department of Pathology

Cardiff University

A thesis submitted for the degree of

Doctor of Philosophy

2008

Abstract

The cell cycle, with its highly conserved features, is a fundamental driver for the temporal control of cell growth and proliferation in tissues - while abnormal control and modulation of the cell cycle are characteristic of cancer cells, particularly in response to therapy. A central theme in cancer biology is to resolve and understand the origin and nature of innate and induced heterogeneity at the cell population level. Cellular heterogeneity - comprising structural, temporal and functional dimensions - is a confounding factor in the analysis of cell population dynamics and has implications at physiological, pathological and therapeutic levels.

There is an exceptional advancement in the applications of imaging and cell tracking technologies dedicated to the area of cytometric research, that demand an integrated bioinformatics environment for high-content data extraction and interrogation. Image-derived cell-based analyses, where time is the quality parameter also demand unique solutions with the aim of enabling image encoding of spatiotemporal cellular events within complex cell populations. The perspective for this thesis is the complex yet poorly understood nature of cancer and the opportunities offered by rapidly evolving cytometric technologies. The research addresses the intellectual aspects of a bioinformatics framework for cellular informatics that encompass integrated data encoding, archiving, mining and analysis tools and methods capable of producing *in silico* cellular fingerprints for the responses of cell populations to perturbing influences. The overall goal is to understand the effects of anti-cancer drugs in complex and potentially heterogeneous neoplastic cellular systems by providing hypothesis testing opportunities.

Cell lineage maps encoded from timelapse microscopy image sequences sit at the core of the proposed bioinformatics infrastructure developed in the current work. Through a number of data mining, analysis and visualisation tools the interactions and relationships within and between lineages have provided dynamic patterns for the modulation of the cell cycle in disease and under stress. The lineage data, accessible through databases implemented during the current study, has provided a rich repository for pharmacodynamic (PD) modelling and validation and has thus laid the foundation for fabricating a comprehensive knowledge base for linking both cellular and molecular behaviour patterns. These provide the foundation for meeting the aspirations of systems biology and drug discovery.

To the inspirational memories of my father.

Acknowledgements

I cannot start this section just by thanking my supervisors as this research did not start like any other PhD research. The work initially stems from a belief rather than a hypothesis which was transpired by the inspirational talk given by Rachel Errington back in Autumn 2004, who later became my supervisor, mentor and friend. I definitely owe her a big thank you for the development of the PhD structure, but also my heartfelt gratitude to her for stimulating the intellectual content; as well as shaping my cellular informatics aspirations and career. I owe another big thank you to Paul Smith (co-supervisor) for his inspirational and mentorship role, who has always challenged my endeavours from the cancer biologist's perspective . I am also indebted to him for his continuous support in every respect – from logistic to intellectual. Both Paul and Rachel cannot even realise how much I have learned from them and the everlasting impact they have made on me. Another big thanks to Richard White, my third supervisor who has always overviewed the informatics aspects of this research and gave me the much needed technical support.

I would like to thank the other colleagues in the wet - lab who made the work possible by acquiring and encoding the data – Janet Fisher, Marie Wiltshire, Lee Campbell, Nuria Marquez, Kerenza Njoh, Sally Chappell and Emeline Furon. Without their comradeship attitude it would not be possible to build this bioinformatics infrastructure from nothing. At times their help was more personal than professional, like Janet lending her laptop to me for writing the thesis at home.

I would like to thank my colleagues at Cardiff University who have lent a hand on various occasions: Ian Brewis for constant support in every aspect, Julie Williams for providing the funding in the initial years, Valentina Moskvina for the statistical analysis, Alex Hardisty for his discussions and inputs and finally David Kipling for my current funding.

I would like to thank the people who have collaborated with this project externally and have provided their support, intellectual inputs, time and efforts. Nick White of Oxford University; Reza Ali, Judith Pérez-Velázquez, Neil Evans and Mike Chappell of Warwick University; Moncia Lupi and Paolo Ubezio of Instituto di Ricerche Farmacologiche "Mario Negri," Milano, Italy; and Patricia Martin and Scott Johnstone of Glasgow Caledonian University. A special thanks is also given to Christopher Kier of Molecular Devices,

USA, an industrial collaborator whose team has responded to requests for software development; this gives my work an additional value as it can be rolled out to a larger user community.

I would also like to convey my thanks to some of my students who have contributed to this research through their research projects, Peter Husemann for co-building ProgeniDB and Christopher Hedley for his Johnson curve analysis.

Finally I would like to thank my mother Husenara for her blessings and my family - Rumana my dearest wife and Zarif my beloved son, without their unconditional support and limitless affection this work would not be possible.

Chapter 1: Introduction

1.1 Bioinformatics – turning biological information into knowledge

With the completion of the human genome project (HGP 2003) and scientific advances in the post-genomic era, the life sciences have witnessed an enormous volume of information generated by both biotechnological research and instrumentation development. Arising from the demand to apply information gathered to form knowledge and understanding for clinical and other benefits - a new interdisciplinary science of bioinformatics has evolved (Hagen 2000). The endeavour started in the early 1980s with the methods of DNA sequencing (Simpson 2001) and now encompasses genomics (Burley 2000; Lockhart and Winzeler 2000; McKusick and Ruddle 1987; Nadkarni 2002), proteomics (Abbott et al. 1998; Dove 1999; Ho et al. 2002; Jensen 2006; Twyman 2004), and in recent years metabolomics (Harrigan and Goodacre 2003; Joyce and Palsson 2006). The advancement has been coupled with a continual development of experimental technology for the acquisition of molecular biology data quantitatively and accurately (Abbott et al. 1998; Bruggeman et al. 2007; Lincoln 2001). In parallel, information technology has also witnessed a major advancement in terms of data management and data access, e.g. the public use of the internet (Castells 2001). These parallel advancements have transformed bioinformatics from a data management technology to a discipline where the ultimate goal is to transform experimentally derived biological information into knowledge (Heidorn et al. 2007) and thus enable the discovery of new biological insights as well as to create a wider perspective from which unifying principles in biology can be discerned.

Until recent years, a significant part of bioinformatics was service-oriented (Foster 2005), focussed towards the common needs of information technologies in large-scale biological data. However, lately the drive towards transforming information into knowledge is prevalent in all areas of bioinformatics (Kanehisa and Brok 2003). For example, in the genomics area where primary databases like 'Entrez Gene' (Maglott *et al.* 2005) archive all the gene-related information, whereas secondary databases, like KEGG (Kanehisa 2002), integrate and cross-reference numerous databases in a multi-species context and fabricate a better understanding about biological function from a genomic perspective. The same paradigm shift is also apparent in proteomics where, for example, secondary databases like BIND (Bader 2003) have been introduced that capture protein function, here defined at the molecular level as the set of other molecules with which a protein interacts or reacts along with the molecular outcome. Metabolomics is yet to keep pace with its other '-omic' counterparts and encompasses primary databases like the Human Metabolome Database (Wishart 2007), which provide a comprehensive curated collection of human metabolite and metabolism information. The list of databases is growing at a fast pace and presently the number of molecular databases is over 1000 (Galperin 2008). These databases, with associated smart mining and analytical tools, e.g. BLAST (Altschul *et al.* 1990) provide both information and knowledge that directly contribute to our understanding of the molecular basis for disease as well as the structural and functional complexity of cellular processes that constitute the organ or organism.

1.2 Gene to organism – issues of scale and complexity

A molecular basis for understanding organism behaviour started with the premise that organisms are assembly of different components which can be described in a hierarchical fashion according to their functionality, size etc. From a size perspective, an organism like a human can be described at many scales, and the lowest level is represented by a defined atomic/molecular description - for example the genome. Thus the genomic level can be assigned to the lowest level, which can define potential within an organism's specific structure and function. From this level upwards everything is the product of causation from genes to cells, organs, systems and whole organism (Dawkins 1976)

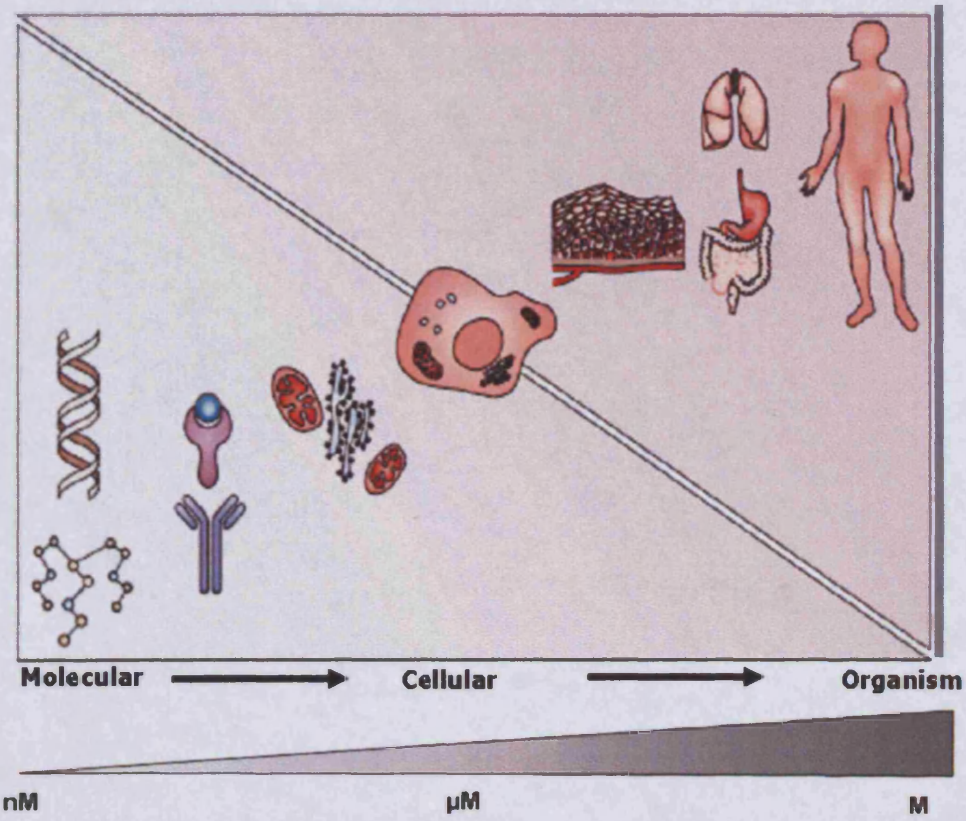


Fig. 1-1 ***Scaling the building blocks of life.*** *The human as an organism is the product of different parts. From DNA or gene, which is at the nanometer scale; proceeding upward to the micrometer scale (cell) and finally the organism at the meter scale. The cell sits at the middle of this scale and acts as a bridge between genotypic constituents with phenotypic outcome (adopted from (Anderson et al. 2001)).*

However this product-based premise could not be equivocally translated in terms of the scale of the complexity. As large scale data and knowledge emerge from genomics, proteomics and other '-omics', it becomes apparent that they are not sufficient for understanding the higher complexity of biological systems (Kanehisa and Brok 2003). The prevalent bottom-up approach (Bruggeman *et al.* 2007) starts at the gene level and moves upwards, and involves 20,000 – 30,000 protein coding genes as elementary information units. The associated genetic networks involving gene sequences and the perception of an increasing number of sense-antisense transcription units and non-coding co-regulatory RNAs (Valet *et al.* 2004) are precursors of millions of different proteins in different functional states leading to a combinatorial number of billions of possible interactions to test (Figeys 2004). Knowledge from these interactions could enable us to understand cellular and organismal phenotypes at the systems level (Wu and Bonner 1981) . However, this understanding demands an unprecedented number of hypotheses, experimental work, mathematical model simulation and assimilation, which are largely beyond the capacity of present technologies (Allen *et al.* 2001).

Again from the experimental point of view, these large scale '-omic' data are not *bona fide* representations of the innate cellular conditions, due to the methods of collection and exclusion of natural complexity and heterogeneity. For example, genomic and proteomic data are usually collected after destruction of cell integrity and cell environment in the original tissue, and the structural and functional parameters measured at these conditions may not reflect the *in situ* condition (Valet 2005b). Moreover the dependence of functionality on the context (such as experimental condition, cell status and environment) at present are mostly ignored (Kanehisa 2000). Last but not least, these data represent a snapshot or static interaction, which indeed is an incomplete view of the dynamic condition where all interactions as well as events are time-dependent. An example in this regard would be the failure of 3D protein structure prediction (Aloy *et al.* 2003) from the known amino acid sequences, despite substantial progress in computing potential and software development and intensive molecular biology research for over 30 years (Valet 2005a). Regardless of the acknowledgement that the bottom-up approach (Stransky et al. 2007), with its paradigm that the molecular basis of knowledge is the key for understanding the disease process and biology at the system level, has a biased and often limited view. For the past decade pharmaceutical industries have introduced a molecular target-based drug discovery approach, where a target is usually a single gene, gene product or molecular mechanism, in which the process of drug discovery begins with identifying the function of a possible therapeutic target and its role in disease (Kerns *et al.* 2003; Knowles and Gromo 2003; Lindsay 2003). This approach is different from the empirical physiology-based approach (Erickson 2003), where compounds are screened and profiled based on the readouts of the amelioration of a disease phenotype in an animal model or cell-based assay. Identification of the drug target and the mechanism of action would follow in later stages of the process by deduction based on the specific pharmacological properties of lead compounds. Even though these two approaches are not mutually exclusive, this paradigm shift not only caused a decrease in the number of new chemical entities (NCE) discovered but more importantly new drugs, found to be pharmacology active at their molecular targets, impart toxicity through other targets at the system level (Sams-Dodd 2005). Two such classic examples are the low-density lipoprotein cholesterol lowering anti-atherosclerosis drug cerivastatin (Lipobay) (Psaty *et al.* 2004) and the anti-inflammatory cyclo-oxygenase 2 (COX2) inhibitors (Melnikova 2005). These experiences may induce a shift of efforts (Schneider 2004) towards the search for drugs effective on distributed targets as for example, salicylic acid acting on various molecular targets simultaneously (Rainsford 2007). The limitations of a bottom-up approach and the widening information gap – how genes and molecules specify the systems behaviour, invoke the strong requirements for top-down views (Anderson et al. 2001): a living system is more than the sum of its parts and it acquires emergent properties that its individual components may not have (Zhang et al. 2002). Explaining these often counterintuitive properties in terms of the underlying components requires the cell to be placed as the irreducible and integrating unit that links molecular information with behavioural information.

1.2.1 Understanding biology at cellular level

According to the cell theory, a cell is the smallest living unit in any organism (Schleiden 1838; Schwann 1839). The modern tenets of cell theory maintain that the cell is the structural and functional unit of all living organisms and is generated from the pre-existing cells by a reductive or non-reductive division, where in each division hereditary information is passed from the mother cell to the daughter cells. According to differentiation status, cells represent the elementary functional units of multicellular organisms, and disease represents molecular alterations that impact upon the integrity and functions of cellular systems determined by both genotype and external or internal influences (Valet 2005a). Single cells thus integrate the structural and functional information from molecular pathways and networks to underpin the often asynchronous population (tissue) behaviour, which in turn generates physiological system function. Thus cells can be viewed as the middle level between molecular and whole organism behaviour, encapsulating all the molecular drivers (i.e. gene, proteins, metabolites and the functional networks) in a minimally bounded system capable of integrating extra cellular influences from neighbouring cells as well as environmental factors and hereditary influences maintained in a pedigree structure. Thus, the cell provides an opportunity for a middle-out-approach (Bray 2003; Brenner 2001; Noble 2002a; Noble

2002b) for dealing with experimental data and intellectual concepts (Noble 2002b). However, even within the cellular level there are layers of complexity that are required to allow for integration of information and knowledge arising from both bottom – molecular and top – tissue/organ levels. Within this thesis three sublayers of complexity are described , to be transferred into an informatics framework.

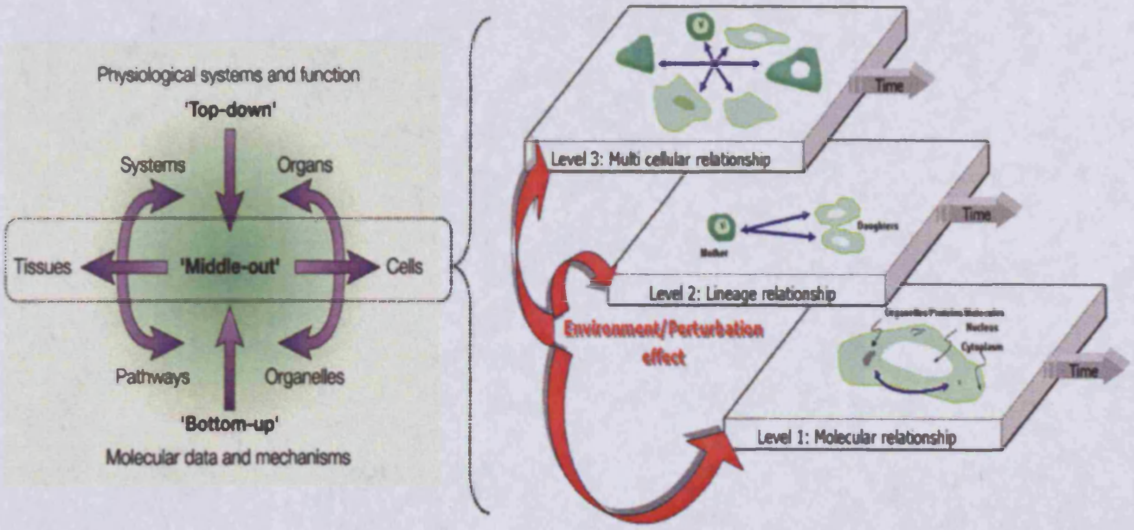

Physiological systems and function
'Top-down'
Systems
Organs
Tissues
'Middle-out'
Cells
Pathways
Organelles
'Bottom-up'
Molecular data and mechanisms
Time
Level 3: Multi cellular relationship
Time
Daughters
Mother
Level 2: Lineage relationship
Environment/Perturbation effect
Organelles/Proteins/Molecules
Nucleus
Cytoplasm
Time
Level 1: Molecular relationship

Fig. 1-2 ***Interweaved levels of bio-complexity.*** *The left diagram schematically shows different approach for understanding biological systems [adopted from (Noble 2002b)] and the right diagram shows the sub-layers that exist within the cellular level that are necessary to assert the functional relationship. This establishs bridge between the molecular level and the tissue or organ level.*

Within the cellular level the first sub-layer (L 1: Molecular relationship) addresses single cell structure, which encapsulates the molecular networks and pathways. Present endeavours for a systems level understanding of biology involve modelling such pathways and networks to gain knowledge about the higher systems or organism behaviour – a 'bottom up approach' (Stransky *et al.* 2007). However this approach does not include the interactions of cells in a time dependent manner – in other words dynamic 'cytome' behaviour. Where the cytome can be defined as cellular systems, subsystems and functional components of the organ and organism (Valet *et al.* 2004). Within these dynamic systems, cellular responses to perturbation or other environmental effects change with time and this time dependent behavioural variation can permit prediction of complex lineage behaviour – the second sub-layer (L 2: Lineage relationship). In experimental terms, a cell lineage reflects the relationship between descendents from a common progenitor that was exposed to a given influence for a discrete period. The behaviour of both the progenitor and the evolving progeny reveals the time-integrated response (e.g. variation of multi-cyclic behaviour) to an influence such as a bioactive drug (i.e. the pharmacodynamic (PD) response). This would therefore have a direct relevance, to how cellular populations, that represent resistant fractions, might be maintained in drug-treated tumour cell populations. The third sub-layer (L 3: Multi cellular relationship) addresses the multi-cellular system (cytome) that illustrates the dynamic interactions between cellular systems or subsystems and environment and provide opportunities to model and predict homogeneous and heterogeneous behaviour of the cytome. Sensitive yet high throughput technology for data acquisition of these multi-dimensional, multi-scalar dynamic data sets depicts an insightful description of living systems (Anderson et al. 2001) and the relationship among these data sets is a prerequisite for our understanding of biology at a systems level including disease processes (Pollok 2005).

1.2.2 The depth and breadth of single cell information

Single cell analysis by image or flow cytometric methods has reached high throughput capacity in recent years (Bullen 2008; Valet 2005b); High-throughput and indeed high-content cell-based screening systems, incorporating elegant reporter assays, have been effectively used to profile drugs based on simple stimulus-response readouts (Bullen 2008; Terstappen et al. 2007). These include high throughput single cell microscopy (Bocsi *et al.* 2004a; Ecker *et al.* 2004a; Ecker *et al.* 2004b; Gerstner *et al.* 2004; Kantor *et al.* 2004; Mittag *et al.* 2005b; Perlman *et al.* 2004b; Schubert 2004) with data reconstituted, to single cell molecular 3D tissue architectures (tissomics) (Ecker and Tarnok 2005; Kriete and Boyce 2005; Schubert 1990, 2004). High throughput flow cytometry (Edwards *et al.* 2004) or flow and image hybrid systems (George *et al.* 2004) as well as chip-based flow systems (Palkova *et al.* 2004; Weston and Hood 2004; Wu *et al.* 2004), cellular genomics (Taylor et al. 2004), cellular proteomics by immunophenotyping (Casasnovas *et al.* 2003; Maynadié *et al.* 2002; Valet *et al.* 2003) and chemical cytometry (Arkhipov *et al.* 2005; Dovichi and Hu 2003; Wu *et al.* 2004) as well as cellular metabolomics (Dovichi and Hu 2003) constitute further facets of recent extensions in molecular cytomics. However the design of current high-throughput instrumentation discards biological heterogeneity, and most assays never contend with dynamic processes. In the absence of detailed kinetic information, simple snap-shot or static high-content-assays provide an over-simplified and often skewed view of the cellular system.

Encoding and organizing cytometric information, especially image cytometry-derived kinetic information, and transforming that into pertinent knowledge within a bioinformatics context is the core theme of the present research. Cell theory informed levels of bio-complexity as outlined in figure 1-2 is the basis for this endeavour where the cell cycle is an underlying and driving force for this complexity due to its ubiquitous and dynamic nature and arguably the most fundamental process for eukaryotic cells (Nurse 2000a). The premise of the current work is that mammalian cell cycle can provide the mechanistic driver (engine) for cellular dynamics and hence underpins the construction and temporal complexity as outlined as levels 1 and 2 in figure 1-2. This approach enables the incorporation of the important characteristics of proliferating cellular systems including: checkpoint controls, alternative cell cycles, asymmetry of division, lineage (multi-cycle) responses, cellular interactions and the evolution of drug resistance (innate and acquired). The ambition is to connect thé nature and probability of cellular responses with the analysis of early molecular decision events – linking origins and outcomes separated over wide timescales. The gap is considerable because of the problems of data acquisition in providing both informative and standardized single cell read outs and the bioinformatics challenges of encoding and interrogating the spatiotemporal cellular perturbations.

By developing a bioinformatics framework, this research aims to provide benefits by contributing to the current understanding of complex cellular dynamics and associated mathematical model building. Models that attempt to fabricate predictive cell response profiles will have use in pre-clinical drug screening, experimental therapeutics and hypothesis-driven research, a common interest shared by a wide range of life scientists (Carnero 2002; Malumbres and Barbacid 2001; Sampath and Plunkett 2001; Walker 2001). The cell cycle has been the subject of intense and varied study over the past 100 years (Nurse 2000b), and investigation of the basic molecular mechanisms is set to continue apace providing a long-term demand for linked bioinformatics solutions (Nurse 2000b).

1.3 Cell cycle – the engine that drives population dynamics

In eukaryotes, the cell cycle involves numerous regulatory proteins that direct the somatic cell through a specific sequence of events culminating in mitosis and the production of two daughter cells while germ cell generation and gamete fusion are modulations of this theme (Smith et al. 2008). The precision with which cell cycle events are executed ensures the survival of living organisms, while loss of this precision increases genomic instability, an important factor in the formation of cancer (Nurse 2000b). Various proteins regulate this progression through different stages of the cell cycle which, from a morphological aspect, can be divided broadly into two phases: interphase (I), during which the cell grows, accumulating nutrients needed for mitosis and duplicating its DNA, and mitosis (M) phase, during which the cell normally divides into two daughter cells.

1.3.1 Phases of the cell cycle

Soon after division each daughter cell begins the interphase of a new cycle, which again divides into subphases. Although these subphases of interphase are not easily distinguishable by morphology, each phase has a distinct set of specialized biochemical processes that prepares the cell for quiescence or a potential cell division event. The first subphase of interphase, which can be mapped from the previous M phase up to the beginning of DNA synthesis, is called G1 (G indicating gap). This phase is marked by synthesis of various enzymes required in for DNA replication in S phase. The duration of G1 is highly variable, even among different cells of the same species (Smith and Martin 1973). The ensuing S phase starts when DNA synthesis commences; when it is complete, all of the chromosomes have been replicated, i.e., each chromosome normally having two (sister) chromatids. Thus, during this phase, the amount of DNA in the cell has effectively doubled. Rates of RNA transcription and protein synthesis are relatively low during this phase. An exception to this is histone production, most of which occurs during the S phase (Nelson et al. 2002a; Wu and Bonner 1981). The duration of S phase is relatively constant among cells of the same species (Ivan and Greulich 1963). The last subphase of interphase is G2, which lasts until the cell enters metaphase. Significant amounts of protein synthesis occur during this phase, mainly involving the production of microtubules, which are required during the process of mitosis. Inhibition of protein synthesis during G2 phase prevents the cell from undergoing mitosis (Stefansson and Brautigan 2007).

After interphase the next phase is M (Mitosis) phase which is again divided into four subphases - prophase, metaphase, anaphase and telophase. During prophase, the replicated chromosomes, each comprising two identical chromatids, are condensed into compact packets and then released to the cytoplasm when the nuclear membrane breaks down. During metaphase and anaphase, the chromosomes are sorted, and each chromatid of a pair moves to opposite sides of the cell. The end of mitosis is marked by a re-formation of a membrane around each set of chromosomes which is designed as telophase. Division of cytoplasm, also known as cytokinesis, generates two daughter cells, each with a 2n complement of genetic material (Lodish et al. 2004).

After mitosis (cell division) both daughter cells again enter to a G1 interphase and from this phase not all "post-mitotic" cells may enter a subsequent S phase by respecting a G1 cell cycle check point, thus providing a non-proliferative fraction of cells in a G0 state. G0 cells may remain quiescent for long periods of time, possibly indefinitely (as is often the case for neurons) particularly following full differentiation. Some cell types in mature organisms, such as parenchymal cells of the liver and kidney, enter the G0 phase semi-permanently and can only be induced to begin dividing again under very specific circumstances; other types, such as epithelial cells, continue to divide throughout an organism's life. According to their location, state and function, cells may also be destined for programmed disposal through apoptosis - a highly regulated process by which an organism eliminates unwanted cells without eliciting an inflammatory response. Apoptosis is involved in many physiological processes including tissue homeostasis, embryonic development, and the immune response (Schwartzman and Cidlowski 1993). The timing and order of cell cycle events are monitored during cell cycle checkpoints that occur at the G1/S phase boundary, in S phase, and during the G2/M phases (Murray and Hunt 1993). These checkpoints ensure that critical events in a particular phase of the cell cycle are completed before a new phase is initiated, thereby preventing the formation of genetically abnormal cells. These checkpoints ensure that critical events in a particular phase of the cell cycle are completed before a new phase is initiated, thereby preventing the formation of genetically abnormal cells (King and Cidlowski 1998).

1.3.2 Regulation of the cell cycle

Because it is essential to identify and eliminate cells proliferating inappropriately, apoptosis and proliferation are tightly coupled, and cell cycle regulators can influence both cell division and cell death (Meikrantz and Schlegel 1995). Two key classes of regulatory molecules, cyclins and cyclin-dependent kinases (CDKs), determine a cell's progress and direction through the cell cycle (Nigg 1995) under the regulatory influence of CDK inhibitory molecules and processes that provide for specific activation and destruction. CDKs are serine/threonine protein kinases, with a wide range of target molecules involved in cell cycle progression, being activated through phosphorylation at specific points in the cell cycle. There are at least seven CDKs in mammalian cells (Pines 1995). The CDKs are critical for progression through the cell cycle because their inactivation prevents mitosis (Devault et al. 1991; Parker and Piwnica-Worms 1992; Van den Heuvel and Harlow 1993). Cyclins form the regulatory subunits and CDKs the catalytic subunits of an activated heterodimeric holoenzyme; cyclins have no catalytic activity and CDKs are inactive in the absence of a partner cyclin. When activated by a bound cyclin, CDKs perform a phosphorylation that activates or inactivates target proteins to orchestrate co-ordinated entry into the next phase of the cell cycle. There are several types of cyclins and most of them bind to a particular type of CDKs and are active at different phases of the cell cycle. However, there are several “orphan” cyclins which have no CDK partner, for example cyclin F is an orphan cyclin which is essential for G2>M transition (Fung and Poon 2005; Karp 2007; Lee and Zaho 2006).

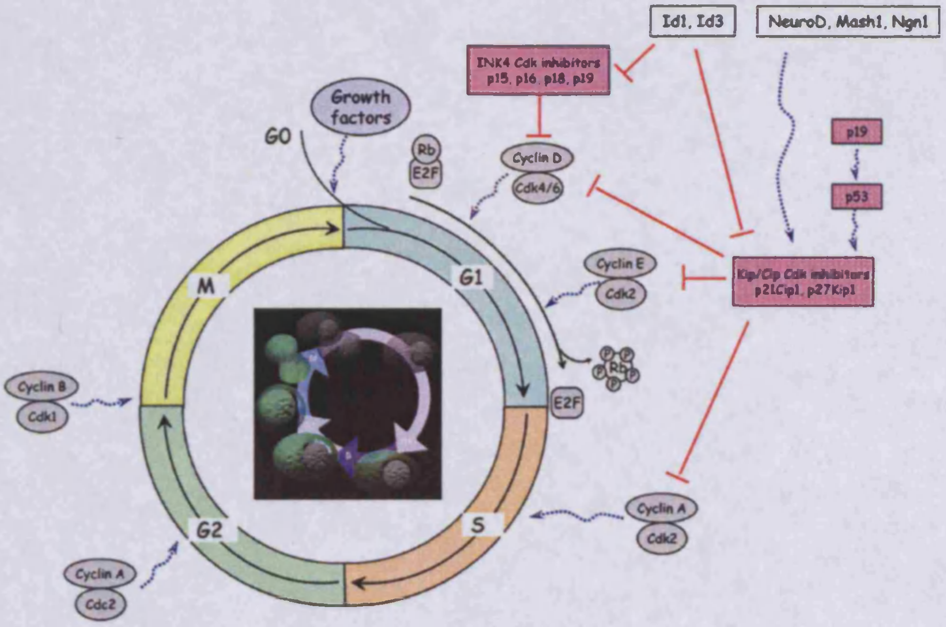


Fig. 1-3 ***The cell cycle and its regulation***. *Schematic representation of the mammalian cell cycle and its regulation. Activating and inhibiting influences are represented by blue arrows and red lines, respectively. In the inset of the schematic representation is a typical expression profile of a fluorescent (eGFP-linked cyclin B1) cell cycle phase marker (CCPM) for G2/M and associated cellular shape changes observed during a cell cycle. At the end of S phase dim cytoplasmic fluorescence is observed and during G2 phase this fluorescence becomes more intense. As the cell enters the early stages of mitosis (prophase) green fluorescent protein (eGFP) translocates to the nucleus. As mitosis proceeds the nuclear membrane dissolves and the cell rounds up and becomes intensely green. Finally, towards the end of mitosis the reporter is degraded so that the two daughter cells are non-fluorescent [adopted from (Abrous et al. 2005)].*

The concentration of some cyclins like cyclin B1 varies in a cyclic fashion during a cell cycle in relation to their production or destruction. When the concentration of a particular cyclin is low it detaches from the corresponding CDKs, and inhibits its kinase activity, probably by causing a protein chain to block the enzymatic site (Bai et al. 1994; Kong et al. 2000). Cyclins and CDKs are valuable markers of cellular proliferation. The kinases are expressed differentially in cells undergoing cell cycle progression, and cyclins show phase-specific expression. These differential expression patterns act as proliferation indices, providing numerical values for subpopulations of cells in different phases of the cell cycle.

Many drugs differentially target phases of the cell cycle and this heterogeneity in cellular responses presents a means of cell cycle-mediated drug resistance. For example, a cytostatic anticancer drug, exerting its effect at specific phases of cell cycle results in a modified and complex cell cycle traverse pattern - delay, arrest and checkpoint breaching. This PD response at the single cell level is governed by complex signal transduction and regulatory pathways, which in turn are determined by the prevailing pharmacokinetic (PK) response – the form and fate of the perturbing agent within the biological system. The relationship between the PK and PD responses are complicated, usually requiring mathematical description, especially when the perturbed biological system expresses discrete events within a heterogeneous cellular population (Chappell et al. 2008). Here a more detailed understanding of cell cycle dynamics and their complexity has been categorized at three different levels and a prerequisite to this is the encoding of the interlinked structural and functional information within a standardized data format.

1.4 Cytometry - an analytical technology

Cytometry embraces all aspects of analytical approaches to the characterization and measurement of cells and cellular constituents for biological, diagnostic, and therapeutic purposes. Cell-based measurements can be used to inform aspects of heterogeneity including the asynchronous timing of specific processes and the spatial changes in participating molecules and structures (Smith 2004). Areas of research and diagnosis include: immunophenotyping (Gerstner et al. 2006; Mittag et al. 2005a; Perfetto et al. 2004), rare cell detection (Bocsi et al. 2004b) and characterization in the case of stem cells (Bou-Gharios et al. 2004; Jovcic et al. 2004; Lovell and Mathur 2004; Rashid et al. 2004) and residual tumour cells (Shen and Price 2006; Steiner et al. 2000; Szaniszlo et al. 2004), tissue analysis (Ecker and Steiner 2004; Gerstner et al. 2004; Megyeri et al. 2005) and drug discovery (Van Osta 2006; Van Osta et al. 2006). The measurement principles used in cytometric systems can be varied (e.g. impedance, dielectrophoresis etc.) but usually employ light. Cytometry-based platforms have offered increasingly sophisticated levels of multiplexing, often based on the principle of light manipulation, can be broadly classified into two groups – *flow and image cytometry* - both of which can be used to supply parameters for use in computational biology to support, for example, both the screening and development of molecular therapeutics (Smith et al. 2007b).

1.4.1 Flow cytometry – high throughput cross sectional data elucidating the cell cycle

Flow cytometry (FC) is a high precision technique for the rapid analysis and even sorting of cells. FC analysis of biological material by the detection of light-absorbing or fluorescent properties of cells or subcellular fractions, such as chromosomes, passing in a narrow stream through a laser beam. Usually the stream is organized to undergo hydrodynamic focusing to allow for optimal presentation of the suspended object in the stream to a focus of the laser beam. Cells can be counted and their laser scatter and fluorescence signals analysed (e.g. intensity, spectral quality and even life-time). A typical 'jet-in-air' system allows for the stream to be broken into droplets with different charges prior to passing through opposed electrostatic plates. In this configuration FC can be used automatically to sort droplets (i.e. fluorescence activated cell sorting; FACS) containing object of interest by deflection at the plates and the sorting of successive droplets of the stream into different fractions depending on the fluorescence emitted by each droplet. The correlated data sets obtained by FC also allow simultaneous multi-parametric analyses of the physical and/or chemical characteristics of single cells flowing through the optical and/or electronic detection apparatus The technique provides statistical accuracy and sensitivity, which is primarily attributed to its large sampling capability (10^5-10^6 cells/ml; >1000 cells per sec) and potential for robotic sample handling. Converting this voluminous multi-parametric data into information and eventually to knowledge is a discipline in its own right and produces a significant challenge (Watson 1991). The major applications of FC measurements include: the analysis of cell cycle, surface and internal epitome detection, apoptosis and cell death detection, functional studies on cellular subcompartments (e.g. mitochondria, lysosomes etc), intracellular signalling pathways (e.g. phosphoprotein expression), analysis of protein location, immunophenotyping (HIV profiling), detection of cellular pathogens (e.g. malaria parasites), cytoskeleton studies and stem cell detection. The measurement of the DNA content of cells was one of the first major quantitative applications of flow cytometry and involves fluorescence detection of nucleic acid content in cells pre-labelled with an appropriate fluorescence tag (e.g. a cell permanent cationic dye such as Hoechst 33342 or DRAQ5 (Smith et al. 2000)) or a combination of labels (e.g. a cyclin reporter and a DNA content dye; Fig 1-4). FC can not only quantify the fluorescence signals but also use these to separate cells of interest from a mixed population based on some pre-selected characteristics (cell cycle phase and integrity) permitting the subsequent analysis of other molecular features.

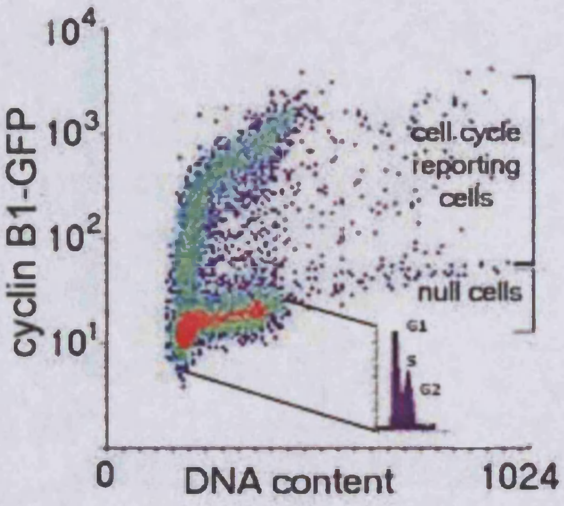


Fig. 1-4 ***Quantifying cell cycle from FC result.*** *The plot here shows a typical FC output where green fluorescence protein (GFP) was tagged with cyclin B1, an endogenous cell cycle marker. Inset shows a histogram representing the proportion of cells present in different phases of the cell cycle.*

In FC, multi-parameter measurements are typically performed to discriminate very specific cell populations to confirm or refute a hypothesis. These analyses provide an essential and cost-effective data source for mathematical modelling particularly in a limited high throughput screen (HTS) mode. However only a fraction of the available information is usually extracted by visual evaluation of multi-parametrically gated histograms or by quantification of marker positive or negative cells (Valet 2005a). This time consuming compartmental approach not only skews results but also lacks the interpretation of interlinked cellular behaviour. Taking repetitive samples of cells undergoing some kinetic process (e.g. drug uptake, drug efflux or enzymatic conversion of a fluorogenic substrate) can inform population kinetics and identify the presence of subpopulation responses. However, in conventional FC measurement it is not possible to track and extract cellular data continuously at the single cell level and therefore correlative or time series analysis on for a given cell cannot be achieved. Moreover the need to gate FC data sets, often subjectively, makes FC data qualitative and non-standardized data (Ubezio and Rossotti 1987). These limitations have implications for

FC as an analytical tool for wider research and clinical use. Recent technological advances have seen the development of laser scanning cytometry, in which cells are scanned on a slide surface and can therefore be revisited at known locations, to allow for kinetic studies. Further, collaborative efforts are underway to establish standardized solutions for representing, collecting, annotating, archiving, analysing and disseminating flow cytometry, data (see: http://flowcyt.sourceforge.net/). Finally a major drawback is that flow cytometry in principle, is unable to analyse cells in their natural environment (e.g. tissues, cell cultures), and the need to reduce samples to single cell suspensions is a complication for adherence-dependent biology or the analysis of cellular interactions and relationships. Consequently other methods have been sought to analyse dynamic events *in situ.*

1.4.2 Timelapse microscopy – potential for continuous single cell tracking

Cellular imaging involves the use of a system or technology capable of visualizing a cell population, single cell or sub cellular structure, applied in combination with image-analysis tools (Lang et al. 2006). Microscopes, in one form or another, constitute a considerable part of imaging technologies and usually generate two dimensional array of information (a digital image) extracted from a particular biological situation. Microscopy can employ different parts of the electromagnetic spectrum for image formation, with two common approaches being optical/visible light-based microscopy and electron microscopy. Both light and electron microscopy involves the diffraction, reflection, or refraction of radiation incident upon the subject of study, and the subsequent collection of radiation signals in order to build up an image. This process may be carried out in wide field mode (for example standard light microscopy and transmission electron microscopy) or by the scanning of a fine beam over the sample (for example confocal microscopy and scanning electron microscopy). Through different microscopic techniques it is possible to acquire images of cellular details at different levels of intricacy, and timelapse microscopy (TLM) is the repeated collection of a single field of view from a microscope at discrete time intervals through which dynamics, for example of cell division, can be captured. TLM enables tracking of single cell events or cellular responses in a population context with a liked time signature (Marquez et al. 2003) and can employ different microscopy modes.

History of TLM

The application of timelapse imaging, to resolve events too complex or fleeting for examination by observation, parallels the development of photography. The pioneer, Eadweard Muybridge is famous for his split-second studies of motion which began in 1872 with an attempt to capture the movement of a galloping horse. Initially, cellular timelapse image sequences were acquired using silver halide–based film, the resultant movies being instrumental, for example, in demonstrating the dramatic behaviour of cell motility (Bajer and Bajer 1972). Over the last 25 years, cell biology has benefited from improvements in electronic imaging technologies that have largely replaced silver-based film recordings. During the 1980s the use of analogue video technology greatly expanded the use of light microscopy as an analytical tool (Inoué 1986; Inoué and Spring 1997; Salmon 1995). Over the last several years, the use of analog video-capture systems has been largely replaced by computer-based digital image capture systems (Inoué and Spring 1997; Sluder and Wolf 2003). With their high quantum efficiency, low-noise characteristics and ease of use, imaging systems for digital microscopy have greatly improved the study and quantification of dynamic cellular behaviour.

Recent advances in imaging technology have been coupled to improvements in photochemistry/photobiology, with the development of sophisticated molecular probes that have allowed the visualization of discrete molecules within living cells (Lippincott-Schwartz et al. 2001; Rieder and Khodjakov 2003). These advances in probes have allowed sophisticated molecular interactions to be studied at the level of the individual cell (Cardullo and Parpura 2003). A major advance in the bio-imaging field has been the development of green fluorescent protein (GFP) (Lippincott-Schwartz and Patterson 2003), which allows tagged proteins to be visualized and imaged. Chimeras made from a gene of interest coupled to GFP — or genetically engineered chromatic variations of GFP — can be readily introduced into cultured cells as well as genetically tractable organisms such as yeast, flies, worms, and fish (Haraguchi 2002; Zhang et al. 2002). More often than not, these GFP fusions retain their native biological activity while becoming fluorescently tagged. Central to the assay is a fluorescent probe that consists of two main components (i) the targeting portion and (ii) the chromophore portion, which presents the signal to be measured. The robustness and dynamic range of the assay is dictated by the efficacy of the ligand-target interaction together with the quantum efficiency of the fluorophore. The choice of fluorescence detection instrument is determined by the nature of the intensity signal to be acquired and the required spatiotemporal resolution. Increasing sophistication in the design and application of biological models as well as the advent of novel fluorescent probes have led to new demands on molecular imaging systems to deliver enhanced sensitivity, reliable quantification, and the ability to resolve multiple simultaneous signals - multiplexed and multispectral imaging (Levenson et al. 2008)

Types of timelapse microscopy

Light microscopy involves passing visible light transmitted through, or reflected from, the sample through a single or multiple lenses to allow a magnified view of the sample (Abramowitz and Davidson 2007). There are multiple types of light microscopy, common approaches being: - bright field, dark field, oblique, phase contrast, differential interference contrast, fluorescence, confocal laser scanning and deconvolution microscopy. The focus of the present research involves the use of phase contrast and fluorescence microscopy and as such only these will be discussed in further detail. Within the time frame of this project it was only possible to cover these two basic approaches although it is acknowledged that other modes have common features and specific advantages or drawbacks.

Phase contrast microscopy

Phase contrast microscopy, a widely used mode for TLM, was developed on the methodology introduced by Zernike in early 20th century (Zernike 1942, 1955). Phase contrast (Goldstein 1982; Inoué 1986; Yamamoto et al. 2003; Zemike 1958) is a technique in which the influence of specimen thickness and refractive index on the phase of light passing through it is used for contrast enhancement by manipulating the phase and amplitude of the un-diffracted light relative to the diffracted light. This mode yields excellent contrast and axial resolution when used with video enhancement (Inoue 1989, 1990; Inoué 1986). Transmission phase offers a probe-less contrast mode providing low resolution but highly informative outputs (e.g. cell shape and cell position) for tracking cell division, cell death and motility (Stephens et al. 2004). This approach has been successfully used in screenings (e.g. in 18-36 well formatted culture plates) to determine single cell cycle traverse, checkpoint breaching in response to drug perturbations (Marquez et al. 2003) and wound closure (Stephens et al. 2004); such a non-perturbing mode can be used for event/time-encoded cell-based assays.

Fluorescence microscopy

Fluorescence microscopy has long been used to capture the details of molecular patterns, distribution and dynamic behaviour at single cell level (Dunn et al. 2004). The attributes, advantages, and uses of fluorescence microscopy are well documented (Agard et al. 1989; Arndt-Jovin et al. 1985; Axelrod 1989a; Axelrod 1989b; Axelrod 1989c; Brakenhoff et al. 1990; Bright et al. 1989; Haugland 1992; Jovin and Amdt-Jovin 1989; Pawley 1989; Taylor and Salmon 1989; Taylor et al. 1986; Tsien 1989; Wang and Taylor 1989). Advances in fluorescent probe design and synthesis (Haugland 1992; Loew 1988; Tsien 1989; Tsien and Waggoner 1990), and molecular biology and protein chemistry (Wang and Taylor 1989), coupled with technological improvements in microscopes and detectors have further enhanced the advantages and extended applications and performance. Molecular tagging using cell cycle (e.g. GFP-cyclin stealth fluorescent reporters (Thomas 2003)) can reveal underlying molecular events and offers a continuous readout for cell cycle and lineage mapping. Intercepting the continuously sampled process by fixing the culture and probes, using an 'in-cell' molecular mapping approach, allows one to obtain a functional and structural fingerprint of linked dynamic data.

1.5 Cellular dynamics viewed through timelapse microscope

TLM has become an important mean to dynamically quantify single cell response to a perturbed situation in a population context (Lang et al. 2006). Multimode microscopy—defined several years ago as the automated combination of multiple modes of light microscopy, including fluorescence, luminescence and transmitted light modes (Farkas et al. 1993) — has emerged as a powerful tool in the dissection of molecular events within living cells. The coupling of multiple channels of fluorescence, whether independent (Plymale et al. 1999) or combined through ratio imaging (Pap et al. 1999), has been applied to a wide range of multimode applications. The resolution, and hence type of event, is determined by the contrast mode which includes phase, differential interference contrast, dark field and fluorescence imaging (White and Errington 2005). The past decade has witnessed an increase in the dimensionality of the cellular information that can be obtained with light microscope methodology (Taylor et al. 2001). The multimode microscope can acquire images at variable rates, at milli-pico second scale for the study of protein-protein and drug-protein interactions. Millisecond resolution is used for the study of ion transportation (e.g. calcium transportation and signalling) while imaging performed at the minute, hour or even day scale is used to track cellular behaviour in a population context (Marquez et al. 2003; White et al. 2005). Transmission/phase offers a probe-less and essentially non-perturbing contrast mode - providing restricted resolution but highly informative outputs on cellular behaviour (e.g. cell shape and cell position) facilitating assays that assess critical global cell PD responses such as the interruption of cell division, induction of cell death and changes in cell motility (Stephens et al. 2004).

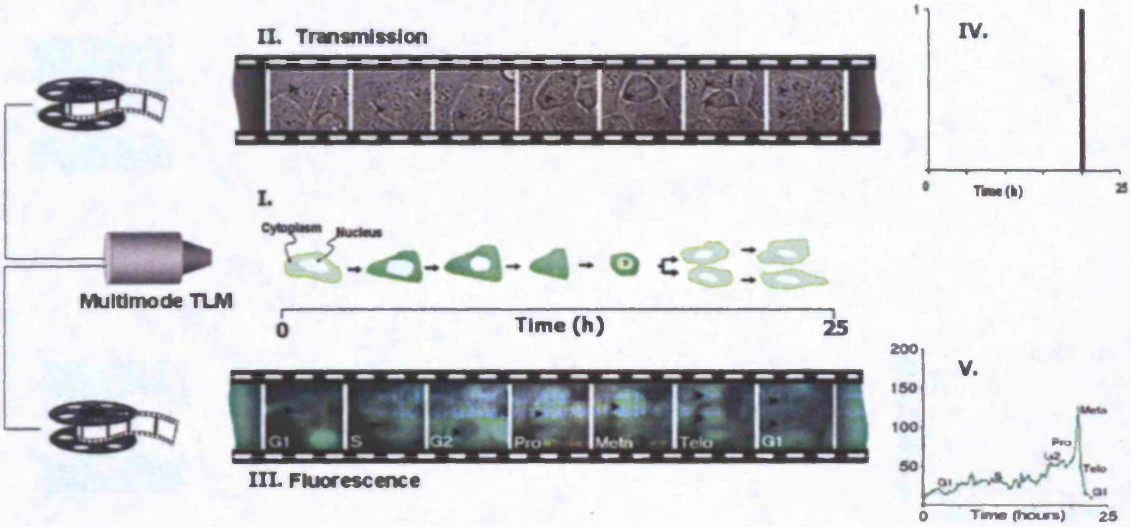


Fig. 1-5 ***Multimode timelapse microscopy capturing cellular dynamics.*** *(I) diagrammatic representation of a cell cycle, where a newly appeared cell at time 0 goes through different phases of the cell cycle and accordingly changes its shape and finally divides into two daughter cells ~20 h. The green colour shade represents the intensity of cyclin B1 tagged with GFP, eGFP-cyclin B1. (II) An actual image sequence acquired via the multimode TLM in transmission phase, the black arrow within the image sequence indexing the cell under tracking. The change of cell shape corresponds to the diagrammatic representation. (III) Simultaneously acquired image sequence via fluorescence phase, the same cell is indexed by arrow but in this mode the resolution is higher and as such the shape of the cell could not be resolved but the tagged protein intensity (eGFP-cyclin B1) can be quantified easily from the image sequence. (IV) An event based plot showing a value of 1 when a cell divides into two daughter cells, based on the shape change visualized in II, number and time of events can be determined. (V) eGFP-cyclin B1 profile extracted from fluorescence image sequence and depicts the intensity at different phases of same cell cycle. From the intensity it is possible to profile the cyclin B1 expression which is an indicator for cell cycling positioning.*

The timelapse approach enables the determination of single cell cycle traverse, delay, arrest and checkpoint breaching in response to perturbations (Giuliano et al. 2005). While the key events of cellular dynamics can be measured by transmission phase, fluorescence phase has been used to measure both the temporal and spatial dynamics of single and multiple proteins within populations of single cells (Taylor et al. 1984). Tagged protein-tracking provides sub-phase information on cell cycle progression, cell-cycle regulator dynamics in parallel with morphological landmarks and DNA content analysis. For example, the application of GFP and imaging techniques to cell cycle analysis has enabled significant advances to be made in understanding the timing of the molecular events that control the cell cycle. Fusing GFP with key cell-cycle control proteins (Arnaud et al. 1998; Huang and Raff 1999; Raff et al. 2002; Weingartner et al. 2001; Zeng et al. 2000) and other cellular components (Kanda et al. 1998; Reits et al. 1997; Tatebe et al. 2001) has been used to study the molecular organization behind the cell cycle. Tracking cells as they respond to pharmacologically active agents using a non-invasive approach provides a means of linking causative events with later outcomes at the molecular level, and forms the basis for molecular response fingerprint or pharmacokinetics (PK) response. Here, multimode microscopy – transmission and fluorescence, has enabled us to visualize and parameterise cellular behaviour at different levels of feature resolution – morphological, molecular and behavioural (e.g. event outcome of division, arrest, delay and death). These parameters collectively start to provide a comprehensive – more holistic - map for the study of cellular dynamics which is yet to be fully exploited due to the hurdles of data management and informatics frameworks.

1.6 Converting images to numbers - data management issue

Advances in imaging technologies, in particular the development of high performance/low noise camera systems, and a parallel increase in computational power, have enhanced the ability to acquire and manage multi-parametric TLM data of increased complexity and quantity (Bullen 2008). Despite these advancements, image analysis has been held back substantially by limitations in the software used to store, process, and analyse such large volumes of information (Goldberg et al. 2005). Current software for microscopy automates image acquisition but fails to provide a robust data format through which these images can be annotated, stored and accessed. The primary reason has been that the derivation of information from images is completely dependent on contextual information that may vary from experiment to experiment (Goldberg et al. 2005) and that invokes a need to couple cellular descriptors with experimental descriptors in order to reveal complete information (metadata) about the dynamics. This is where cytomic data are fundamentally different other 'omic' data i.e. genomic data, which is independent of experimental variation. For example the form of a 'correct' DNA sequence is not dependent on the type of sequencer used and this feature is not included in sequence annotation. Moreover cellular information extracted from images, until recent years, were purely 'expert driven' and qualitative, demanding intensive effort. With the advent of commercial image analysis software (MetaMorph, ImagePro, MATLAB, GE Health Care) along with open source software like ImageMagic, transformation and to some extent obliteration the 'qualitative' aspect is achieved but the labour intensive aspect of image analysis still persists due to the lack of automation. So the key challenge remaining to-date is the development of image analysis algorithms that automatically extract information at single cell level (Price et al. 2002). Timelapse image analysis has been described as one of the greatest remaining challenges in screening (Echeverri and Perrimon 2006; Eggert and Mitchison 2006) and this field of biological science is also described as "very much in its infancy" (Murphy et al. 2005) and "lags behind the adoption of high-throughput imaging technology" (Perlman et al. 2004b). Addressing this bottleneck caused by the qualitative and labour intensive nature of the technology, different open source projects like CellProfiler (Carpenter et al. 2006) and ImageJ-NIH (Collins 2007) have emerged, focusing on the development of analysis packages and algorithms that can extract image-derived cellular information. Such as morphological information, from fields of cells and can address a variety of biological questions quantitatively, including standard assays (for example, cell count, size, per-cell protein levels) and complex morphological assays (for example, cell/organelle shape or subcellular patterns of DNA or protein staining) in high content and through-put modes (Carpenter et al. 2006).

Cell-based assays are conveniently prepared in multi-well culture plate formats, such as 96-well and 384-well plates, for high-throughput screening to facilitate the study of responses of a population of cells under different chemical, genetic, or radiation perturbations. However, in the absence of integrated solutions to image data management, it has become standard practice to migrate large amounts of data through multiple file formats as different analysis or visualization methods are employed. Once analysis is carried out, the results are usually exported to a spreadsheet program like Microsoft Excel for further calculations or graphing. Due to the lack of proper data management i.e. connectivity among different stages (image acquisition to data analysis), it has become almost impossible to coherently dissect or query all the elements of the data management environment. Moreover the data model used in any imaging system varies from site to site, depending on the local experimental and acquisition system. Finally, even within a particular site, data models change over time as new acquisition systems, imaging technologies, or even new assays are developed. The development and application of new imaging techniques and analytic tools will only accelerate, but the requirement for coherent data management and adaptability of the data model remain unsolved (Goldberg et al. 2005).

From this evolving demand for a new approach to microscopic image and image derived data management, OME Open Microscopy Environment (OME) (Goldberg et al. 2005) (also see: www.openmicroscopy.org) was established where the primary goal was to enable the automatic analysis, modelling, and mining of large image sets with reference to specific biological hypotheses. OME aims to manage and store the original image along with the metadata that specify the context or meaning of that image. Some metadata are devoted to describing the optics of the microscope, some to the experimental setup and sample, and some to information derived by analysis. Finally, OME aims to provide a flexible mechanism for incorporating new and existing image analysis routines and storing the output of those routines in a self-consistent and accessible manner. It thus forms an image informatics infrastructure where it would be possible to manipulate and share image data as readily as genomic data; from a genomic perspective, this approach can be compared to the MIAME approach (Brazma et al. 2001) which aimed to standardize microarray experimental descriptors and annotations. OME has defined general image information to be five dimensional (5D): coordinates X and Y, focal point Z, time T and wavelength or channel C. Through XML schema image information is tagged and thus supports systematic and quantitative image analysis as well as formulating primary standards for image data. Through OME-XML schema metadata and through OME-TIFF, image files are stored in a relational database, which can be shared via server protocols.

The bottleneck of handling and analyzing large number of image data has prevented image analysis from being performed in a high content fashion. High content screening (HCS) is a powerful approach for disease diagnosis and prognosis, drug target validation, and compound lead selection (Zhou and Wong 2006). It has recently emerged as a promising solution to improve the quality of decision making in drug development (Bullen 2008). The challenge lies in how to convert all the images showing functions and interactions of macromolecules in live cells and tissues into quantitative values that can be analysed statistically (Zhou and Wong 2006). Bearing this in mind, in recent years a joint effort between CellProfiler and OME has been launched to implement a complete open-source infrastructure for organizing and analyzing images (Swedlow et al. 2003). Timely inception of such infrastructure and next generation HCS machines from instrument developers (GE Healthcare, Molecular Devices and Chipman Technologies) with longterm live cell culture modules, is set to make a significant contribution to cytometry based high content screening (HCS).

1.7 High content screening – scaling up cell based assays

HCS has introduced a step forward to the current method of TLM similar to the advancement of automated DNA sequencing over manual sequencing methods. This has been accomplished by automating the major aspects of the imaging process, including analysing the huge numbers of arrayed cells that could be tested with a wide range of experimental treatments rapidly and without extensive human interaction. Automation of image acquisition, image processing, image analysis, image archiving, and image visualization has made it possible to prepare large numbers of microplates, placed in a stacker on the HCS instrument and with operator walk-away while the plates are processed by the system. This has permitted an accelerated approach to the process of producing data through to creating new knowledge from a massive number of cells in a matter of one day. HCS has the potential to fundamentally change the process of doing large-scale cell biology in basic biomedical research and drug discovery (Taylor 2006). HCS has made large-scale cell biology a tractable approach to drug discovery by generating functional genomic information through the automated measurement of the temporal and spatial activities of genes, proteins, and other constituents in living cells (Abraham et al. 2004; Giuliano 2003; Giuliano et al. 1997). The foundation of HCS involves the strategic combination of instrumentation, imaging algorithms, reagents, and data visualization, archiving, and mining software to dissect the interrelationships between cellular processes and the effects chemical compounds, including potential therapeutic candidates, have on them (Giuliano 2003; Minguez et al. 2002; Wipf et al. 2000). It is important to note that HCS targets involve not only the direct site of drug interaction but also the multiple physiological processes that are invariably affected by drug activity (Giuliano et al. 2004). HCS assays in which multiple parameters are not only measured within single cells using multiple reagents and morphometrics, but the relationship of these parameter values are calculated, analysed, and interpreted on a cell-by-cell basis (Taylor 2006). Exploiting this multiplexed cellular information has benefited the process of drug discovery by facilitating the early decision making on drug targets, lead selection, and late-stage attribution (Zhou and Wong 2006). Additionally it has also elevated our understanding of the complex biochemical and molecular processes, occurring in time and space, that dictate cell function and the complex behavioural responses of cells to natural environmental changes or experimental treatments – a systems level understanding. By facilitating early decision making HCS is set to ease the bottlenecks in the early stages of drug discovery process which indeed of great importance for pharmaceuticals and from a research perspective the system level understanding of cellular behaviour.

A current trend in systems biology is the reverse engineering of networks to model gene regulatory or protein–protein interactions with a subsequent extraction of basic principles for biological organization and complex disease phenotypes (Schadt and Lum 2006). The ability of *in silico* representations to predict how a system in a particular state may react and adjust to perturbations has made systems biology an attractive component of basic research, drug development, and predictive medicine. However, computational systems biology are not adequately developed in dealing with the spatiotemporal properties of cells and multi-cellular architectures (Loging et al. 2007). Indeed, attempts to integrate and interconnect various levels of biological organizations, such as genes, proteins, cells, and tissues, are in their infancy (Kriete 2005). A recent study has used a timelapse video approach to allow for data linkage in studying the spatiotemporal dynamics of social amoeba (Dictyostelium) cell populations comprising more than 2,000 mutant clones from a large mutagenesis collection. The dataset generated allows one to search and retrieve movies on a gene-by-gene and phenotype-by-phenotype basis (Sawai et al. 2007). Efforts have been made to utilize the cellular information to understand single cell-based response evaluation to drug treatment (Conrad et al.

2004), pattern recognition of localized protein distributions that improve currently available ontologies (Boland et al. 1998), as well as analysis of changes of sub-cellular phenotypes due to systematic RNA interference (RNAi) (Perlman et al. 2004a).

Discarding the rich spatiotemporal context of cells within a complex tissue and assuming that each cell is a separate entity is a clear simplification in attempting to describe the systems behaviour of cells. At a tissue or organ level, a cell cannot be designated as an independent entity, since the behaviour of 'a cell' at 'a time point' is the result of many factors - environmental or experimental variability its predecessor had endured, number of generations that have elapsed since the last experimental or environmental perturbation, age or generation variation relative to surrounding cells and cell cycle phase variability comparative to its siblings. These factors govern the current behaviour of 'a cell' and impart asymmetry to an event outcome. In the somatic cell context, not all cells in the same cell lineage behave identically, i.e. do not divide or die at the same rate especially in disease states or perturbed situations.

Recently the MitoCheck project group (EMBL, Heidelberg) have demonstrated the impact of timelapse microscopy in a pilot automated platform assaying cell division (mitosis) and chromosome segregation to provide a time resolved *phenoprint* of mitotic gene knockdown (Neumann et al. 2006). They have demonstrated that phenotypic classification would be misinterpreted in corresponding endpoint assays, thus data on event patterns such as delayed versus and arrested is the basis for determining and understanding the subtleties through which occult cell cycle pattern become apparent. Moreover it is not only the cell cycle subtleties but also the event outcome that is also asymmetric and needs to be addressed in a relationship context. For example, an outcome event of two identical daughter cells may be opposite to each other (e.g. one daughter cell die while other daughter cell divide), and the asymmetry of events for somatic cells cannot be overlooked as random processes as these two levels of asymmetry ultimately lead to subpopulation heterogeneity. Indeed at some point during the development of an organism unicellular/unipotent and mitotically active cells acquired an ability to undergo an asymmetric division. Through this special type of cell division, these cells could divide to generate two different progeny or to self-renew and at the same time generate a progeny that is committed to become a cell different from the mother cell (Gaziova and Bhat 2007). The relationship of cells in a tissue construct and in an organ system needs to be exploited to understand the population dynamics, and again it is the cell cycle that remains the predominant driving force for this dynamic. Exploiting the molecular and spatiotemporal information within a context where relationships among all participating cells is recognised, lays the foundation for illustrating a truer reflection of cellular dynamics at a systems level.

1.8 Understanding cellular heterogeneity – an analytical and modelling approach

The importance of spatiotemporal information is not only limited to systems biology or drug discovery but also is essential for understanding biological processes such as growth (Palaniappan et al. 2004a), tissue repair (Farooqui and Fenteany 2005), differentiation and metastatic potential (Ronot et al. 2000), and chemotaxis (Dufour et al. 2005). To understand how apoptosis is induced by anti-mitotic drugs that vary in their ability to capture cells in successive mitosis, or how a subpopulation of cancer cells evolve resistance to an anticancer drug by evading cell cycle directed toxicity, requires a metadata level understanding of the influence of perturbing agents on cell cycle progression. Kinetic cell-based assays derived from TLM, where time is the quality parameter, demand solutions enabling image encoding and interrogation of cellular behaviour in a population context (Marquez et al. 2003). Cellular kinetic measurements provide a route to revealing important time windows at cellular level to study the mechanism of action of individual agents and their response pathways and thus establish more precise, quantitative, and multi-parametric characterization of cell cycle mechanisms under different perturbed conditions (Lang et al. 2006). Given the role of the mammalian cell cycle in defining a proliferation response to a perturbing agent, many studies required in vitro synchronization to make coherent sense of population-based assays, and demand a novel data format that encapsulates the features of cell-cell heterogeneity and time-dependent events while maintaining inter cell relationship.

Another approach for understanding cell cycle dynamics is through mathematical modelling, since mathematical modelling of biological process has two distinctive advantages – firstly, parameter estimation and optimisation which is a prelude for developing any model for revealing in-detail understanding and measurement of different occult features which otherwise would be difficult to quantify from experimental data alone. Secondly, through simulation, different scenarios and the consequental outcome can be visualized within fraction of time to that of experimental duration. In a cell cycle context although the knowledge of the biochemistry and the physical processes of the proteins that regulate the cell cycle is fairly recent, mathematical models of the cell cycle can be traced back to as early as the 1970s (Hastings et al. 1977; Tyson 1974/75; Tyson and Sachsenmaier 1978) and even with recent mathematical models (Aguda 1999; Aguda and Tang 1999; Chen et al. 2000; Qu et al. 2003; Tyson 2002) the complexity of cell cycle dynamics could not be comprehensively simulated (Csika´sz-Nagy et al. 2006). A recent review (Clyde et al. 2006) suggests that the deficiency of experimental data is the main reason for the bottleneck and moreover the data need to be of "new kind" comprising higher temporal resolution and permissive for simultaneous multi-parametric quantitative comparisons. Such data need to embrace multiple factors: (i) the data need to embrace the relationship among cells within a system and thus can be used to develop mathematical models that depict complexity and cellular heterogeneity. (ii) these model outputs along with the experimental data should contribute to elevate our understanding of PK and PD response. (iii) the data can be segmented in different levels from single to sub population to whole population level, (iv) the data can also be segmented in relation to different time domains, e.g. in analysing the behaviour for a particular time period or for particular generation of cells. (v) the` data can be easily accessed and shared among researchers and modellers alike. (vi) the data can be shared with other encoded data originating from variety of timelapse experiments (e.g. fluorescence, phase contrast) and other cytometric instruments e.g. flow cytometry. Retaining such multi-parametric complexity, heterogeneity within a relationship context invokes new data format and cell lineage is deemed as the most plausible approach in this regard, as cellular behaviour encoded through lineage format can encapsulate the important features like event, time etc. while maintaining the inter cell relationship. Moreover during data mining lineages can be segmented according to different levels of complexity and time windows.

1.9 Cell lineage – embracing the complexity and heterogeneity

The study of cell lineages has been, and remains, of importance in developmental biology (Stern and Fraser 2001). The human body is made-up of over 100 trillion cells and understanding which lineage relationships might be informative for disease processes, among these vast numbers of cells is a fundamental challenge for developmental biology and other branches of biology (Alvarez-Buylla et al. 2001;

Anderson et al. 2001; Ardavin et al. 2001; Clarke and Tickle 1999; Dor et al. 2004; Kim and Shibata 2002; Noctor et al. 2001; Stern and Fraser 2001) and medicine (Bernards and Weinberg 2002; Hope et al. 2004; Tang et al. 2003; Weigelt et al. 2003; Yamamoto et al. 2003). Studies with cell lineage began with Whitman's (Whitman 1878, 1887) description of cleavage patterns in leech embryos in the 1870s, and continued with descriptions of lineages in many invertebrate animals, including nematodes, sea urchins, and ascidians. Studies of cell lineages have been critical in our understanding of how cell fates are specified in development and how fates are correlated with cell division patterns (Chisholm and Hodgkin 1989). A full cell lineage of an organism is the sequence of cell divisions leading from a zygote to each differentiated cell, during development. The exact topology, cellular phenotypes and distribution of cell fates in a cell lineage—what Wood (Wood 1999) termed the 'cell lineage hieroglyphics' — encodes information about the sequence of molecular and cellular events that generated it (e.g. the activation and repression of particular genes, or the secretion of morphogens). Cell lineages have been most comprehensively described for *Caenorhabditis elegans* to elucidate developmental mechanisms and nematode evolution (Braun et al. 2003; Fitch and Emmons 1995; Sommer et al. 1994; Sternberg and Horvitz 1981, 1982; Vancoppenolle et al. 2000; Wiegner and Schierenberg 1998).

Apart from development biology, the potential of cell lineage on elucidating the complex interplay of cellular heterogeneity has rarely been exploited. Even though work on cell lineages in the somatic cell context started over two decades ago (Potel et al. 1979; Stywester and Dennis 1980), it saw little progress until recently. Multiple reasons can be attributed to this trend - underdeveloped image acquisition instrumentations along with labour intensive cell tracking and qualitative data extraction procedure are thought to be the rate limiting factors (Taylor 2006). Over the past two decades, considerable technological advancement of light microscopy has occurred (Bullen 2008; Taylor et al. 2001) and as mentioned earlier one of the challenges remaining to-date is algorithms that can efficiently track cells in image sequences (Price et al. 2002) and also subsequently recognize and distinguish different cellular phenotypes automatically, (Conrad et al. 2004; Roques and Murphy 2002). Cell/object tracking particularly in an automated manner remains a challenge that hinders the transformation of a cell lineage approach to a high content scale. Different object tracking methods like - centroid method, Gaussian fit method, correlation method, sum-of-absolute differences method, and interpolation method were used (Zhou and Wong 2006) in this context but were not successful primarily due to three reasons – first, tracking cells after the bifurcation point (when the cell under investigation divides into two daughter cells), the second is tracking cells in an interactive cell-cell environment, finally low resolution image sequences (phase contrast) with ill-defined cell edge definition (Swedlow et al. 2003). Considering these current strength and limitations, studies have attempted and successfully produced cell lineages encoded from timelapse image sequences (Chu et al. 2004; Endlich et al. 2000; Forrester et al. 2000; Forrester et al. 1999; Prieur-Carrillo et al. 2003) illustrating the viability and proliferation of uni-nucleated and multinucleated giant cells formed after X-irradiation or apoptotic-induction post-irradiation in p21 gene knock-out cell lines. However, the lineage data format used in these experiments is exclusive for those experiments and cannot be translated or transposed to other experimental situations and is such a cul-de-sac approach.

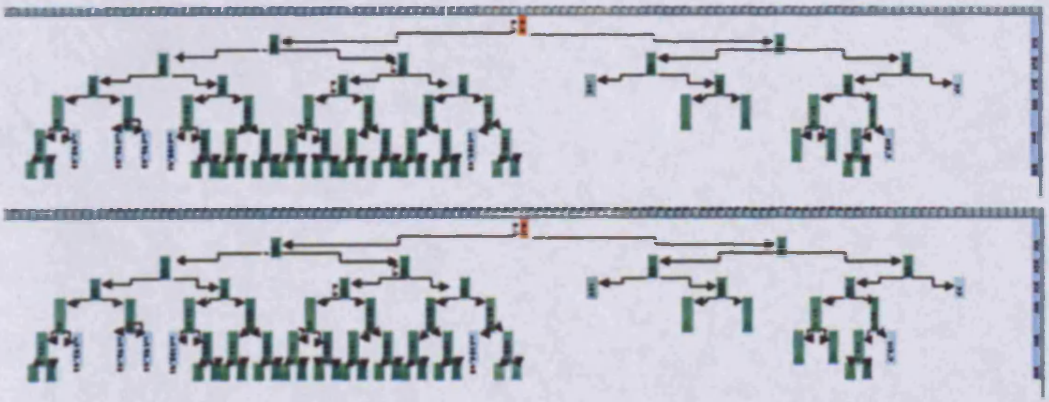


Fig. 1-6 Screen shot of earlier attempts to encode image derived cell lineage data in Microsoft Excel. Individual cells were drawn and data was manually written. The cells or the nodes were then connected by manually drawn lines. This approach was both time consuming, subject to error and was not inappropriate for hypothesis-driven data mining.

Reviewing the chronological perspectives of encoding cellular information in a relationship context, the cell lineage was selected as the most appropriate solution for achieving the "new data" format, since the cell lineage resembles the *in vivo* cellular dynamics. If cellular dynamics are monitored for a specified time duration under certain conditions i.e. tumour or wound healing, one would expect that at the start of the experiment there would be a few or a small number of cells and with the passage of time the cell number starts to increase as a cell proliferates along with asynchronous division which ultimately leads to heterogeneous populations. This time-integrated cell behaviour (e.g. asymmetry in inter-generation cell division time or cell death) can only be successfully depicted via a cell lineage where each node of the lineage represents and stores event related information - division, death etc., while the inter-node link stores life span information (from the start of interphase to the end of metaphase). Finally linking this event and life span (cell cycle) information with experimental or environmental information should permit comprehensive links at various levels and thus establish the desirable multi-parametric relationship in cellular context. Once established this relationship will support metadata analysis, the absence of which restricts current lineage analysis to simple time-oriented analysis rather than a time-integrated relationship analysis.

1.10 Hypothesis and objectives

Acknowledging the impossibility of defining a 'universal data format' that could encompass all timelapse experimental variability and the unavailability of efficient automated cell tracking algorithm/software - the hypothesis is that encoding spatiotemporal cell kinetics data in a lineage format provides a pragmatic route to determining cellular dynamics at molecular, single cell, subpopulation and population levels. These encoded alphanumeric kinetic data integrate multi-scalar events that comprise innate and induced population heterogeneity in dynamic cellular systems and hypothesis-driven interrogation of these data at the metadata level opens a route to revealing the nature and time frames for the modulation of the cell cycle in disease and under perturbed conditions, and forms the foundation for developing mathematical models of cell cycle dynamics. In order to explore this hypothesis certain objectives must be addressed which are inter-dependent and can be broadly grouped into three stages:

1.10.1 Development stage

The objectives during this stage concern formulation and development of a novel data format that encompasses the cellular and experimental heterogeneity in a lineage context. Once formulated the next objective will be to develop encoding tools that can encode cell lineages from TLM within the formulated data format. The final objective during this stage is establishing databases that can archive the encoded data and provide subsequent access to the data, so that a researcher can mine the lineage data for hypothesis driven investigations. As a whole this stage aims to develop an infrastructure within which lineage data encoding, archiving and mining would be possible.

1.10.2 Validation stage

The objective at this stage is to validate the robustness and usefulness of such an infrastructure, especially the data format and accessibility aspects. Common data analysis results, such as cell proliferation rate will be compared between results obtained from encoded data and conventional event counting method (Marquez et al. 2003). Inter cellular relationship – the most critical aspect of the data format will be validated by producing results that illustrate the relationship at different levels (sub-population, experimental conditions etc.). The results should enable a new insight into cellular dynamics and in this way the usefulness of such a data format can be validated. Accessibility to the database also requires validation and this can be achieved by performing typical and novel queries to the database and analysing the mined data. Web accessibility to the database is also an objective in order to validate the usefulness of such database in the public domain.

1.10.3 Evaluation stage

The objective at this stage is to evaluate and explore the effectiveness of this infrastructure, especially the data format in different experimental and biological scenarios. Understanding cellular dynamics at a metadata level is pertinent to a wide spectrum of biological processes and diseases, and cancer remains at the forefront in this regard. The lineage data will be subjected to various analyses to understand the origin of resistance in context of the widely used anti-cancer drug Topotecan (TPT) (Kollmannsberger et al. 1999). The primary reason for selecting TPT was due to the pre-existing knowledge of this topoisomerase inhibitors class of compounds (Pommier 2006) and local research interest. However the overall objective is to elucidate the method of investigating the dose-dependent cellular responses to this phase specific agent. Another anti-mitotic drug Taxol ®, which has a different mechanism of action, will also be investigated to illustrate the scope of the data format.

As stated earlier, mathematical models simulating cell cycle dynamics generally lack the support of experimental data. So, at this stage the objective is to explore what benefits such encoded data could provide to mathematical models. This objective will be explored and evaluated with mathematical/engineering research groups in a collaborative manner. The final and most intellectual challenging objective of this research is to explore the capability of integrating this image-based encoded lineage data with flow cytometric data. Since both image and flow cytometry attempt to elucidate the same cellular dynamics but in different dimensions, integrating such cross platform data could enhance our understanding of cellular dynamics.

While fulfilling these specific objectives, the overall aim to be addressed is the issue of 'generic applicability' - a decisive factor governing the future development and impact of this research. The infrastructure and the data format is required to accommodate not only a wide range of experimental scenarios but also biological processes and diseases like cancer, wound healing and senescence.

1.11 Bioinformatics Infrastructure -beneficiaries

The aim is to develop a prototype bioinformatics infrastructure that demonstrates both the reality of implementing the proposed hypothesis along with the prospect for further development. The infrastructure should have the informatics components that ease the burden of data encoding, data archiving, data sharing and consequently provide bioinformatics attributes such as hypothesis driven data mining and analytical algorithms that would make such an endeavour acceptable not only to cell biologists but also to mathematical modellers.

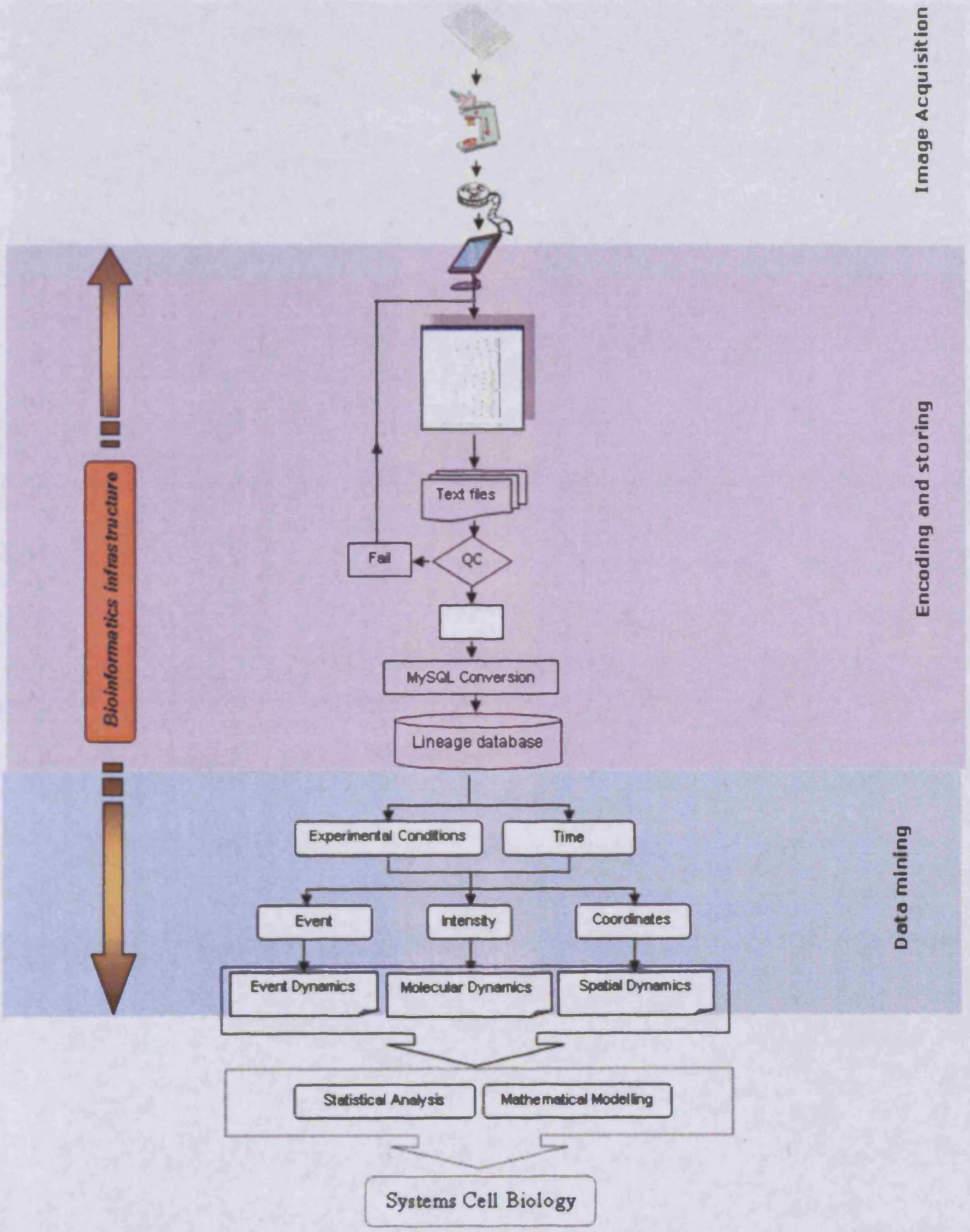


Fig. 1-7 ***Diagrammatic overview of the proposed bioinformatics infrastructure.***

The infrastructure that resulted from discussions with the local users groups and collaborators could be perceived as a construct of three consecutive layers (see figure 1-7). The first layer is image acquisition and arguably does not fall within the bioinformatics remit, but it is important to outline the versatility for experimental and instrument parameters and their incorporation within the data format. The second layer is the core of the infrastructure and includes tools for image viewing and lineage encoding; also included in this layer are the databases that archive the encoded data. The third and bottom layer includes data mining tools and algorithms interfacing with these databases which perform hypothesis-driven data mining. The aim is to explore the lineage data within the sphere of experimental variability and the time domain. The segmented spatiotemporal data acquired through these mining tools will be subjected to typical statistical analysis and compared with previously published results. Access to data will also be made available to the public via web access and especially to the mathematical modeller for exploring and exploiting the effectiveness of such encoded data for modelling cellular behaviour - an indicator for future development and prevalence of such infrastructure.

1.11.1 Element analysis

From the overview of the infrastructure it is clear that a novel data format lies at the core of this infrastructure along with lineage encoding tools, database, and mining tools/algorithms. However to formalize this development an element analysis is required which will outline the key elements that are essential to develop such infrastructure. Considering a typical timelapse experimental situation where image acquisition has already been completed by conventional means, the first informatics tool required would be an image viewer through which timelapse image sequences can be viewed. Once the image is viewed, the cells within the image sequence need to be tracked and image derived information (intensity, coordinates) extracted from the image. Consequently lineage-encoding tools will be required to organize these image derived data in a lineage format, additionally the encoding tools can provide a graphical representation of the lineage in order to orient users in the time domain. The database is essential to store the encoded lineage data while the final element is data mining and visualization tools/algorithms, through which knowledge will be gained.

1.11.2 Specification evaluation

Based on the elements outlined in the previous section a detailed specification needs to be established describing the specific needs of each element.

Image Viewer

i. The image viewer is required to play a wide rage of image files acquired from different acquisition instrumentations.

ii. Since not all image sequences will be of good quality, a wide range of capabilities needs to be in place to improve the image quality.

iii. Image viewer should have the option to play the image sequence in a forward as well as a backward direction.

iv. A wide range of cellular information e.g. intensity, coordinate etc needs to be extracted from any part of the image by positioning a pointer with a mouse.

v. The flexibility of using user-defined tags that can be incorporated with the extracted information is required.

vi. Extracted information along with the users defined tags needs to be parsed to the encoding tool. If the encoding tool is separate from the image viewer, then a bilateral dynamic data exchange (DDE) link needs to be in place through which data can be parsed on both directions. Moreover both image viewer and the encoding tool should have the bilateral executable right on different functionality, so that one software package can run different parts of the other software along with parameter exchange.

vii. Given the coordinates and time, the image viewer should find the specified cell from any part of the image sequence.

Cell tracking

i. Cell tracking is the prelude for encoding cell lineages, and therefore an ideal image viewer should have an automated functionality to track single cell within image sequences and consequently extract information. Moreover, the tracking algorithm also needs to have the capability to recognize different events (mitosis, death) commonly required in cell based assays.

ii. The prime challenge of automated tracking is that most are specialized for single object tracking, however for cell assays the complication arises when the cell divides into two daughter cells. Bearing this in mind the tracking algorithm should be able to track both daughter cells after the bifurcation point.

iii. The amount of supervision needs to be minimized without compromising the quality of the data encoded, moreover it should be robust enough to encompass a wide range of image quality.

iv. When a cellular event occurs, the tracking algorithm should clearly distinguish the start and end point of that particular event.

Graphical view of lineage

i. Since encoding will be performed on a single cell basis, the process of evolving a lineage from a single cell needs to be drawn in real time fashion so that users get orientated with the dynamic process.

ii. Users need to have editing control over the lineage encoding process- delete, add any part or complete lineage.

iii. Options should be in place in the encoding tool, so that when users click any point on the lineage the related point in the image sequence can be viewed within the image viewer.

iv. Selecting the associated encoded information can redraw a previously encoded lineage.

Data format

i. A cell lineage is a set of cells that derive from a single cell often termed as the progenitor cell. The data format should be as such so that each lineage and the associated cells can be indexed uniquely. Moreover each cell and lineage should be interlinked so the relationship can be established at lineage as well as individual cell level.

ii. Data format should include cellular as well as experimental information. At the experimental level the data format should possess the flexibility to embrace different experimental scenarios and at the cellular level should include some basic yet vital information e.g. intensity, coordinates.

iii. All possible event outcome (e.g. mitosis, death) needs to be addressed by the data format along with the capability to mark the start and end of each event.

iv. Data for a single lineage should be primary stored in a text file or XML file, with the ability to be later transferred into a relational database.

v. Once saved the written data can be changed by the encoding tools if editing on the lineage is done.

Data archiving, mining and visualizing

i. All encoded data saved in text or xml flat file format should be primarily achieved in folder structure.

ii. These files will be subjected to quality control to ensure that cellular and experimental descriptors are correctly encoded.

iii. Hypothesis driven data mining will primarily be performed on these quality checked text/XML files. This will give several advantages – first new mining tools can be developed complementing novel ideas regarding data mining and through this process certain mining process will be identified as important and applicable to a wide range of experimental scenarios. Second, analyzing data locally will give the particular advantage of data security, an important aspect for pharmaceutical industries. Third, text/XML file formats will give the flexibility to adopt new mining/analysis tools and new visualization techniques.

iv. Once the mining, analysis and visualization processes are established, the mining technique will use a web accessible database with a relevant interface to provide public access to the encoded data along with ability to perform hypothesis driven data mining.

1.11.3 Domain analysis

The specification set forward invokes not only a complete image acquisition and analysis suite but also the functionality to develop a lineage construction ability along with data warehousing capability. In other words a complete infrastructure was specified which obviously was not available with any of the existing imaging technology. CellProfiler (Carpenter et al. 2006) a felexible and high-throughput image analysis software was useful for image based assays like cell count, cell size determination etc. It was also applicable for morphological assays like cell/organelle shape determination, quatification of subcellular patterns of DNA or protein staining. All these phenotypic or morphological assays could be peformed in an automated fashion sutiable for high-througput technology. However like other commercial software, CellProfiler did not have the analytical components that could be utilized for lineage contstruction from the quantified data, more importantly CellProfiler was more suitable for high content satic images rather than image sequences and thus was not suitable for this particular project. However it is to be imphasized that the open source nature and the modular structure of CellProfiler were in agreement with the outlined specifications and thus can be deemed as a model structure for this project.

A recent comparative study (O'Mahonya et al. 2005) of the four widely used image analysis software packages - NIH-Image, IP Lab, Image Pro+ and MetaMorph in the assessment of the adhesion of micro-organisms to mucosal epithelium using confocal laser scanning microscopy indicated that MetaMorph had particular merit. The 'journaling' feature, through which users can programme and easily build functionality for image, has made MetaMorph an outstanding informatics tools to develop bioinformatics infrastructure. Moreover in MetaMorph users can use their own defined tags and can attach that information along side the image derived information. From a specification point of view MetaMorph meets most of the 'Image Viewer' section's specifications and to some extent the ability to track cells within images with basic capability. However for MetaMorph and other image analysis software outlined, functionality of the software ends with transporting image derived information to popular spreadsheets packages and users perform all the analysis on those spreadsheet data. This does not address the bioinformatics challenges to organize the data in a particular data format i.e. lineage format as specified in this research and consequently does not provide hypothesis driven data mining. The only software that meets part of the specification outlined is Simi BioCell (see http://www.simi.com/en/products/biocell/index.html), a commercial software designed to develop cell lineages from image sequence. With Simi BioCell it is possible to capture and track the entire evolution of an embryo and to document it objectively. Starting with single cells, the capabilities of Simi BioCell go as far as studying the entire cellular evolution of a complete, complex organism. This commercial software is an excellent encoding tool with an embedded image viewer that supports wide range of image file format. However the tracking of the cell needs manual intervention and the data format by which the cellular information gets encoded is specific and applicable to stem cell research (Braun et al. 2003). Like most of the commercial software, the data format cannot be modified according to user specification and finally on the encoded data it is not possible to have a metadata search capability and as such hypothesis driven data mining is not possible.

After careful review of the software packages and their suitability in context to the specifications, it was found that MetaMorph was the most suitable software package to be used and the infrastructure i.e. tools, databases, analysis processes can be developed in association with MetaMorph. The flexibility that MetaMorph offers through journaling and the tag incorporation feature has made this a unique candidate for developing this type of infrastructure. Additionally MetaMorph being acclaimed as one of the most popular image acquisition and analysis software and was already in use within the laboratory at Cardiff where the research was carried out.

It is important to mention here that these specifications are formulated through a process of feedback and evaluation process, which involves Cardiff centric cell biologists as well as collaborators at Warwick University. The design of the infrastructure (see figure 1-7) is the final product of series of discussion sessions with cell biologists, computer scientists and mathematical modellers, same holds true for the specifications discussed earlier. Even though both the infrastructure and the specifications are neither optimal nor complete, it is envisaged that these specifications and the infrastructure suffices the basic requirements of the local user groups and collaborators and forms the foundation for future development.

1.12 Aim

The excitement and opportunities that digital imaging has introduced in recent years to cytometry, has reformed the process of transforming images to numbers. Management, quantification and understanding these numbers in a biological context remains a challenge and needs to be addressed by a wide range of expertise - cell biologist to mathematician. Incorporating such a wide spectrum of expertise requires a 'common platform', which in one hand converts data into information and on the other hand provides access and investigative methods, to gain knowledge from the information. Bioinformatics has proven to be successful in establishing such platforms in other '-omic' domains e.g. NCBI (See http://www.ncbi.nlm.nih.gov/) or EBI (See http://www.ebi.ac.uk/) and the cytomics context efforts are underway e.g. OME (Goldberg et al. 2005) , CellProfiler (Carpenter et al. 2006), BioImage (see http://www.bioimage.org/) to name a few. Based on the hypothesis and associated objectives, this research aims to establish a bioinformatics infrastructure that has the components for future expansion.

Chapter 2: ProgeniTRAK – converting images into numbers

All the timelapse experiments and actual encoding using ProgeniTRAK were performed by Drs Nuria Marquez, Lee Campbell, Janet Fisher, Marie Wiltshire (Department of Pathology, Cardiff University).

2.1 Introduction

Building a bioinformatics infrastructure that encompasses the specifications outlined in the previous chapter and particularly meeting the challenges of incorporating existing commercial image processing packages e.g. MetaMorph (MM), had invoked an in-detailed scrutiny of timelapse microscopy derived data to determine both advantages and limitations. Currently TLM presents clear shortcomings for image interpretation and data management, furthermore the current availability of informatics resources is severely limited for "image" metadata handling (Goldberg 2005) particularly when deciphering basic cellular behaviour and cell relationship at different levels is concerned. The core challenge for formulating an informatics infrastructure that addresses the data management and metadata level interpretation is the design and implementation of a novel data format. The operational challenge is to develop encoding tools that extract data from image sequences and configure the image derived data into a predefined structure consisting of a meaningful language and vocabulary (designated by consensus from the a community of biologists). In short, TLM provides morphological information tracking of single cells, which marks linked key events and cellular responses, where the event resolution depends on the microscopy contrast mode and the spatiotemporal windows of data collection; importantly phase-contrast transmission offers a probe-less and non-perturbing microscopy mode providing outputs on cell behaviour (e.g. cell shape and position), changes in these two basic features facilitate assays describing critical global cell responses such as cell division and cell motility (Stephens et al. 2004). These and other outputs along with contextual data enables further interpretation of cell cycle checkpoints activities, induction of cell death or cell arrest due to acute exposures to a perturbing agent (e.g. such as those induced by anti-cancer drugs) with proliferation consequences for individual cells. Perturbation of a cellular system with continuous or bolus exposure to anti-cancer agents or other drugs provides the capability to convert these basic cell-assays into a quantified pharmacodynamic (PD) response. The process of encoding and encapsulating the detailed cellular response derived from transmission phase TLM within a lineage map is the focus of this chapter. While the fundamental philosophy of a cell lineage map has been described previously, this chapter illustrates the implementation of a lineage approach within an informatics infrastructure and in context of phase contrast microscopy. ProgeniTRAK is the outcome of this endeavour. Bearing in mind the generic nature of the data format as outlined in the specification, the primary focus area of ProgeniTRAK was implementing lineage map of human cancer cell-based assays, later it was used in experiments that addressed p53 gene-knockout, wound healing and mouse primary cell analysis.

2.2 Informatics illustration of a timelapse microscopic experiment

The data format that lies at the core of the informatics infrastructure should embrace descriptors that are essential to describe a typical TLM experiment and at the same time encapsulate the tracked cell behaviour while maintaining relationship within cells. As a result two levels can be identifies for the data format – (i) experimental level and (ii) single cell level, and in combination they form the elements of the overall data format.

2.2.1 Experiment level descriptors

TLM is a widely used technique and covers a diverse range of research and experimental requirements. Formulating a standardized vocabulary that includes such diversity is a huge undertaking and requires large investment and research team like that underpinning the Open Microscopy Environment project (see http://www.openmicroscopy.org/index.html). For the current research the aim was to incorporate the basic features of a typical TLM experiment that are widely used across in common experiments, while always maintaining the view that the approach could be bolted onto or embedded in the future in to environments such as the OME (see chapter 7 for further discussion). A typical instrument used for timelapse microscopy is shown (figure 2-1). Timelapse experiments can be viewed as a combination of interrelated entities and through minimalist approach, these entities and their relationship can be described in a hierarchical fashion (see figure 2-2). At the experiment level, at the top of this hierarchy is the experiment entity and at the bottom is progenitor cell entity. Each entity has its own attributes and relationship with its higher and lower level entity. For example each experiment (or screen) has date, operator name, gassing conditions and other attributes that describe the instrument settings; each experiment also employs culture vessels with multiple wells (ranging from 6 - 96) these elements having unique attributes viz. drug, dose, cell type etc. Again each well can comprise multiple fields of view (FV), another entity with a single attribute – field dimensions. The last entity at this experimental level is the identification and location of each progenitor cell, - the original cell or cells present in the first frame of the image sequence. The progenitor cell therefore also has a few attributes like X, Y coordinates, and possibly the phase of the cell cycle (see later definitions).

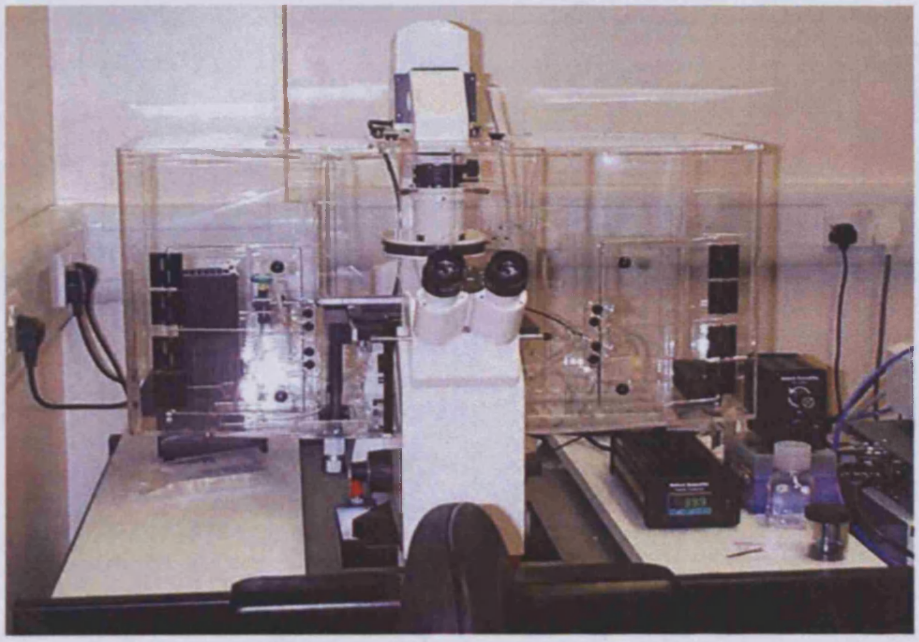


Fig. 2-1 Typical timelapse microscope instrument used to capture typical longterm (48-120 hour) sequences. The instrument consists of a basic inverted microscope with an incubator system to keep cells at 37 ºC and gassed at normoxic conditions (5% CO2). The transmission lamp is fronted by a shutter for periodic image capture (5-20 minute time intervals). Inset: An automated addressable x,y,z stage holding a multi-well plate for undertaking screening experiments. It is these basic instruments and experimental descriptors that are incorporated into ProgeniTRAK.

The above is not a complete list of attributes for any of the entities but provides the basic framework to formulate a nomenclature for identifying each progenitor cells uniquely (see figure 2-2 for the complete attribute map). When these attributes are appended following the hierarchical structure – a tag can be generated which identifies each progenitor cell uniquely. These attributes are selected manually by the user through a series of selection processes which at the end generates the tag, the implication of which in contributing further to the infrastructure will be discussed later in this chapter. The sequential selection process through a bespoke designed graphical users interface (GUI) ensures a relatively error free encoding of the tag which not only identifies all progenitor cells uniquely but is also generic enough to embrace a wide range of TLM experiments.

2.2.2 Cell level descriptors

The lowest and final level of the hierarchical structure is at the single cell level. As each progenitor cell divides a progeny map evolves to form the basic lineage structure. From a biological as well as an informatics perspective, each progenitor cell is also considered as a single cell and thus has the same attributes to that of a single cell – name, coordinates, intensity etc. The cell level attributes serve multiple purposes – primarily as an aid to identify the node of a single cell uniquely within a lineage and subsequently establishes the relationships among cells within a lineage. Secondly the attributes encapsulate the time-dependent birth and behaviour features which later can be analyzed for hypothesis driven data mining.

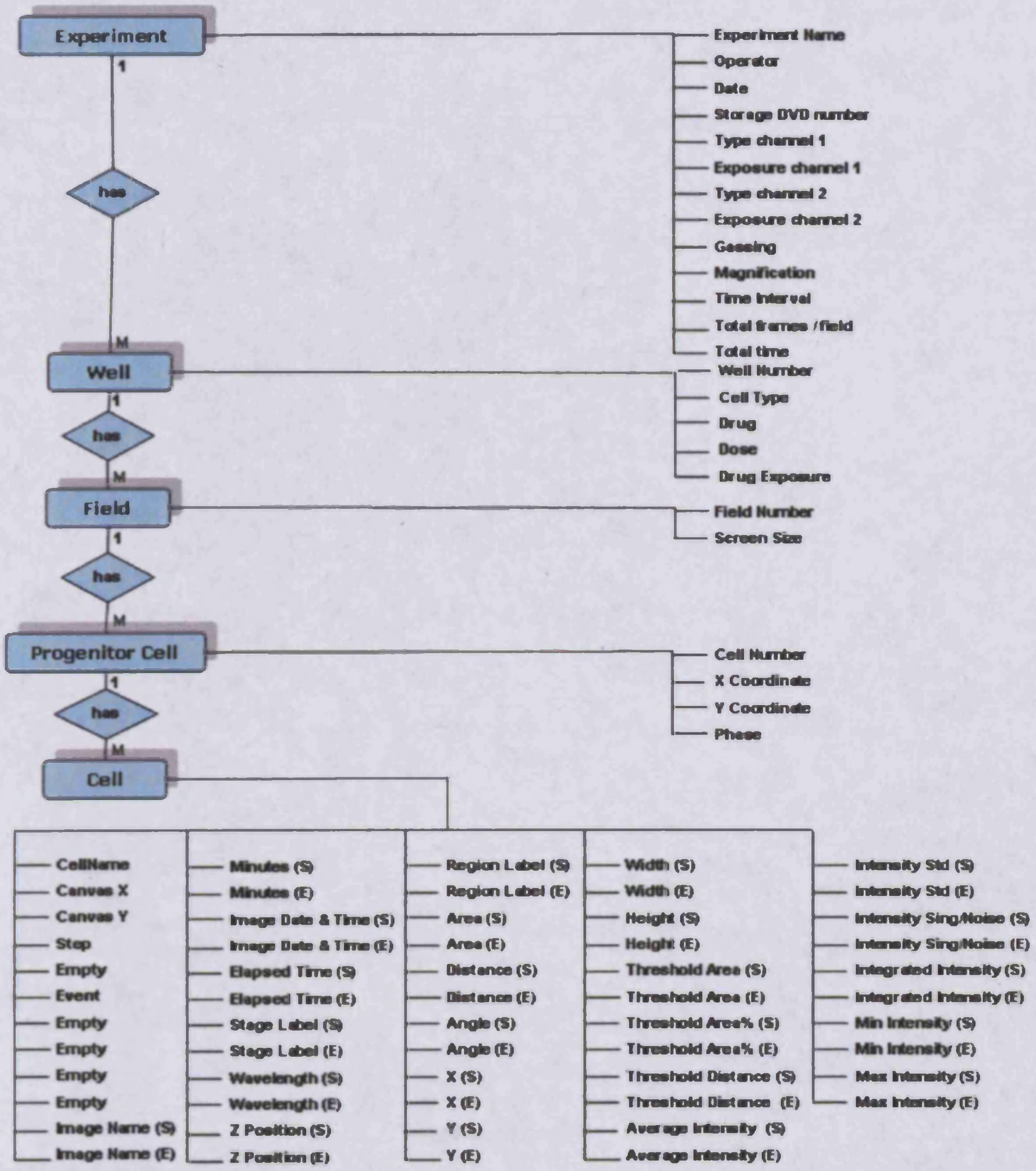


Fig. 2-2 Schematic representation of different entities and attributes of a typical timelapse experiment.

These cell level attributes are directly encoded from the TLM via MetaMorph image analysis software as the PERL script of ProgeniTRAK (see Appendix III) is dynamically linked to MetaMorph. ProgeniTRAK organises these data into a correct order and associates additional tags like cell name, event type etc. to give context to the image derived data. These experimental and cell level attributes forms the basis of encoding the lineage. There are 23 attributes at the experimental level that define the lineage tag and this tag was used as the name of the lineage text file. In the specification it was mentioned that each lineage was required to be saved in a tab delimited text file where each text file represents a separate lineage. This facilitates easy access and mining as well as later conversion to a MySQL database.

Table 2-1 23 parameters that define the progenitor cell as well as forming the text file name.

Experiment Name
Operator
Date
Storage DVD number
Type channel 1
Exposure channel 1
Type channel 2
Exposure channel 2
Gassing
Magnification
Time Interval
Total frames / field
Total time
Well Number
Cell Line
Drug
Concentration
Dose
Field Number
Cell Number
Coordinate X
Coordinate Y
Phase of cell cycle

Exemplar text file name:

LeeLongTerm2_Lee_12052005_DVD561_Transmission_10_GFP_100_5CO2_10_15_452_6780_C1_U2OS_TPT_10UM_BBM60t0_7_20_287_469_U

Single cell (including the progenitor cell) level attributes were encoded within the text file where each text file contains rows of tab delimited data fields and further each single cell constitutes a row and each row consists of 58 data points or attributes as illustrated in figure 2-2. In ProgeniTRAK data were encoded when a major event like cell division, death etc. occurred and for each event data were encoded at the start (S) and end (E). At each instance 24 attributes were extracted (for details see Appendix I) and ProgeniTRAK appended these 24 x 2 attributes or data points into a single row and added additional attributes e.g. canvas coordinate, step etc and standardized vocabularies to give a context to these 48 data points. Most importantly within each lineage or text file, each cell was given a unique name through a nomenclature - the progenitor cell is named as 'B' and if this cell divides into two daughter cells then they are named as 'BN' and 'BS' respectively, where 'N' refers to north and 'S' to south daughter. For a re-fused or polyploidy outcome the designation is 'BE'. If three or four daughters were produced they are named as 'BN' 'BE' 'BS' or 'BN' 'BU' 'BL' 'BS' respectively. It is important to note that even though 'N', 'S', 'E' etc. directional letters were added to the suffix of the cell name, these did not indicate spatial orientation or direction, rather used solely to index each cell within a lineage uniquely.

2.3 The encoding infrastructure

By reviewing the specifications set forward in the introduction chapter and analyzing the data format, it was envisaged that the infrastructure can be broadly divided into two sectors - (i) an image sector which deals with image archiving, viewing and extracting information and (ii) an encoding sector which connecting with the image component via the dynamic data exchange (DDE) link provided by MM and organizing the image derived data to the predefined format as well as actually drawing the physical lineage onto a canvas. In the domain analysis, the usefulness of using MM in this context was justified, currently images are stored in a typical folder structure (each folder archives *n* number of image sequences as files), moreover MM archives 'journal' files in a designated folder which can be accessed, modified and executed by associated user defined button in MM. While MM constitutes the image sector of this infrastructure, in the encoding sector a novel encoding tool was required along with folders to which the experimental descriptors were archived and provided the archive for the encoded lineages as text files. Bearing this in mind a novel encoding tool- ProgeniTRAK was developed, written in PERL 5.8 this encoding tool was designed to parameterize phase-contrast timelapse microscopy image sequences. PERL was considered as the best programming language for writing ProgeniTRAK simply because PERL provided easy reading and writing capabilities with text, MS Excel and the MM Log file (where MM parses the image derived data through the DDE link), moreover PERL through its TK module (a widget development toolkit) can produced an effective GUI providing a dynamic for reproducing the lineages.

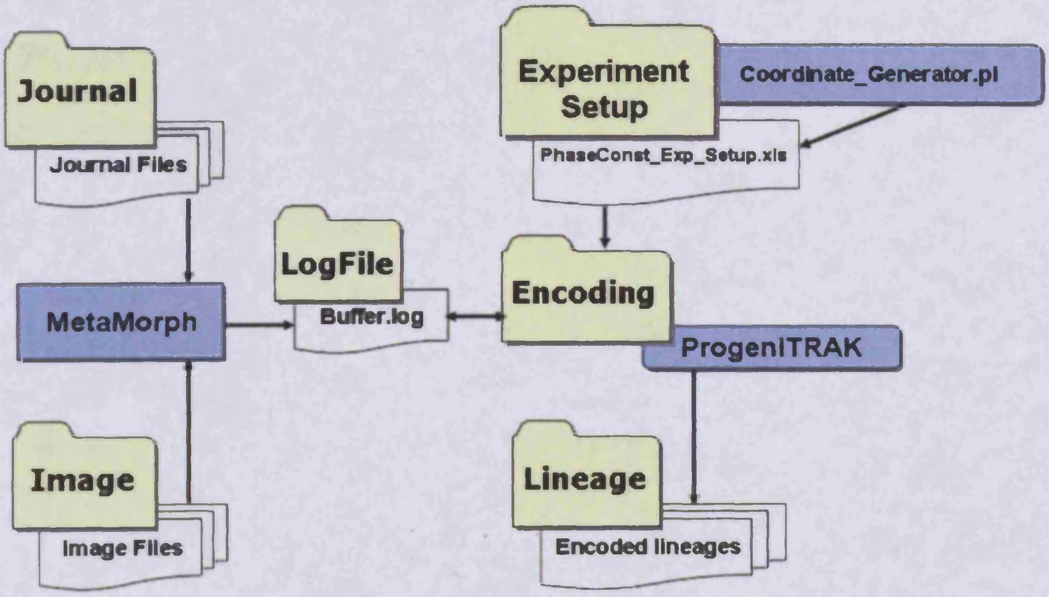


Fig. 2-3 ***Folder structure of encoding environment.*** *Blue represents software while yellow and white represent folders and files respectively.*

Figure 2-3 illustrates the different folders in two designated sectors of the infrastructure where the 'LogFile' folder provides the link between these two sectors. The left side of the figure comprises the image sector consisting of the 'Journal' and 'Image' folder along with the software MM. The right side comprises the image sector with 'Experiment

Setup', 'Encoding' and 'Lineage' folders along with the encoding software ProgeniTRAK and a small PERL script ('Coordinate_Generator.pl') as the main executable component. The 'LogFile' folder, which acts as the link between these two sectors, has a one-way link, implying that ProgeniTRAK can not execute or parse parameters to any of the journal files of MM. Conversely the MM Journal files can execute any of the functionality of ProgeniTRAK as well as parse image derived data via the DDE link. A brief overview of each folder is given bellow.

2.3.1 The image folder

In this folder image files were stored mostly in the proprietary format of MM (*.stk format) however any image sequence is acceptable. Each image file when acquired and archived in this folder was given a three parameters name format - ExpName_WellNum_FieldNum, so that from the file name the relevant image sequence file can be found during the uploading process.

2.3.2 The journal folder

In this folder journal files of MM were stored. These journal files were designed through a process of an iterative development specified by the users and can be executed by the button provided in the menu bar of MM. The process of encoding is directed by a series of buttons in MM as follows:

Image upload

The first button is to start the encoding process. This button opens a file upload dialog box through which user can select the image sequence file (field of view) they are interested of. The image files are kept in the Image folder. This button also actives the 'Regional Measurement' feature of MM that lists the parameters for the data that need to be extracted from the image sequence.

Cell locator

The button helps to locate a cell within an image sequence. Given the frame number and coordinate the journal executed through this button can index a cell within the image by an arrow. This journal is valuable to find the bifurcation points of the lineage.

Log data

The button and linked journal executes the parsing of all the regional measurement parameters to the log file via the DDE link.

Erase data

This button erases the last row of data from the log file.

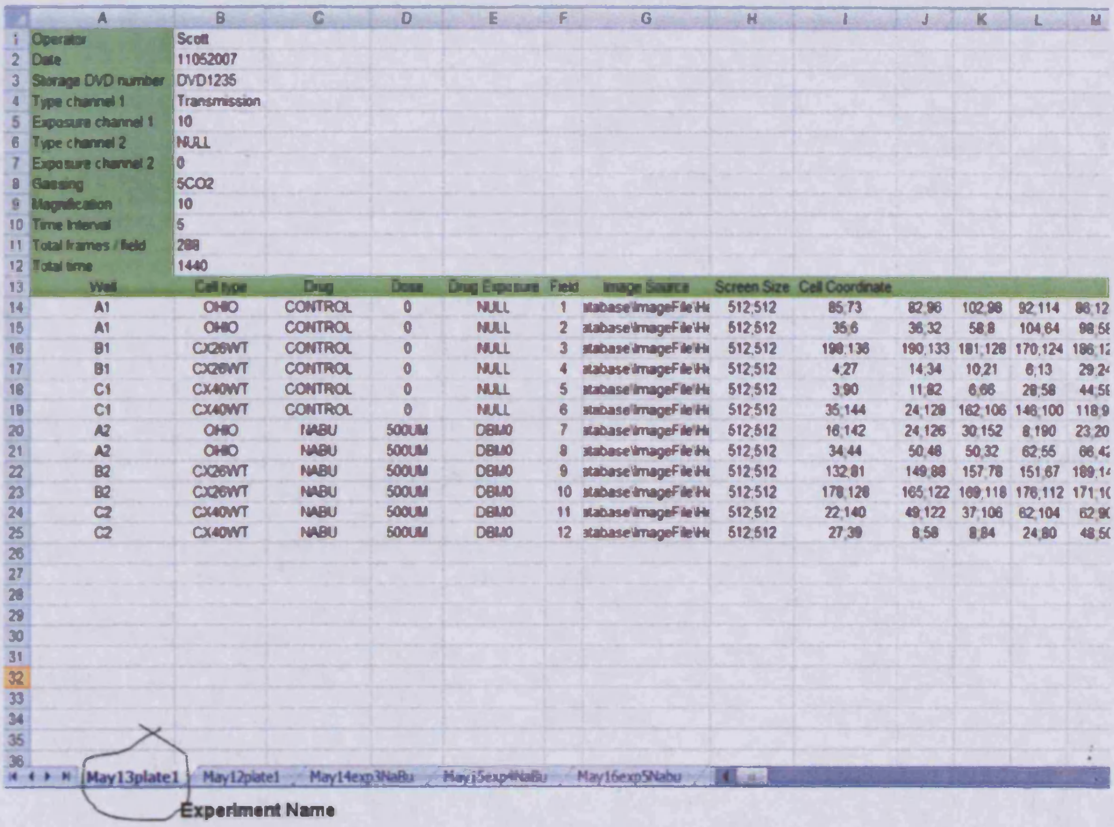
LogFile Folder

This folder holds only the Buffer.log file where MM writes the image derived data dynamically as each region of interest is places over the image and clicked.

2.3.3 The experimental Setup Folder

The folder contains a single Excel file where all experimental descriptors were semi-automatically written by the user, routinely the user completes this record at the time of conducting the experiment and acts as a substantial record of the experiment as well as providing the basis for ProgeniTRAK. This Excel file contains the attributes of experiment, well, field of view (FV) and progenitor cells as outlined in figure 2-4 and each sheet of the Excel file represents an experiment and as such the name of the sheet is renamed according to the name of the experiment. Within each sheet the first twelve rows store the values of the 12 attributes (except experiment name) of experiment. From the thirteenth (14th) row each row represents a Well, i.e. Well attributes . The first five columns of each row stores the values for the five attributes of a single Well and 6th, 7th and 8th column stores the - field number, linked to the image sequence of the FV and field size respectively. From the 9th column onwards each Excel cell stores the coordinates of the progenitor cell and the length of the row is variable depending upon the number of progenitor cell present in the first frame of the FV under consideration. Some parts of the information are encoded manually while other parts are encoded semi-automatically, for example in order to write the coordinates of the progenitor cells, the appropriate image file was opened in MM and a Region of Interest (ROI) was placed on the nucleus of each progenitor cell at the first frame of the image sequence and when clicked, the coordinates of the nucleus were parsed through a DDE link to a log file and a small PERL script was written to read the coordinates from the log file and writes them directly into the Excel file.

This Excel file was named as 'PhaseConst_Exp_Setup.xls' and acts as a 'digital laboratory notebook' that stores all experimental details and since each sheet represents a single experiment, one file is sufficient for storing all experiments of a laboratory.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Operator	Scott											
2	Date	11052007											
3	Storage DVD number	DVD1235											
4	Type channel 1	Transmission											
5	Exposure channel 1	10											
6	Type channel 2	NULL											
7	Exposure channel 2	0											
8	Gassing	5CO2											
9	Magnification	10											
10	Time Interval	5											
11	Total frames / field	288											
12	Total time	1440											
13	Well	Cell type	Drug	Dose	Drug Exposure	Field	Image Source	Screen Size	Cell Coordinate				
14	A1	OHIO	CONTROL	0	NULL	1	atabase\imageFile\H	512,512	85,73	82,96	102,98	92,114	86,12
15	A1	OHIO	CONTROL	0	NULL	2	atabase\imageFile\H	512,512	35,6	36,32	58,8	104,64	98,5
16	B1	CX26WT	CONTROL	0	NULL	3	atabase\imageFile\H	512,512	198,136	190,133	181,128	170,124	186,1
17	B1	CX26WT	CONTROL	0	NULL	4	atabase\imageFile\H	512,512	4,27	14,34	10,21	6,13	29,2
18	C1	CX40WT	CONTROL	0	NULL	5	atabase\imageFile\H	512,512	3,90	11,82	6,66	28,58	44,5
19	C1	CX40WT	CONTROL	0	NULL	6	atabase\imageFile\H	512,512	35,144	24,128	162,106	146,100	118,9
20	A2	OHIO	NABU	500UM	DBM0	7	atabase\imageFile\H	512,512	16,142	24,126	30,152	8,190	23,20
21	A2	OHIO	NABU	500UM	DBM0	8	atabase\imageFile\H	512,512	34,44	50,48	50,32	62,55	66,4
22	B2	CX26WT	NABU	500UM	DBM0	9	atabase\imageFile\H	512,512	132,81	149,88	157,78	151,67	189,1
23	B2	CX26WT	NABU	500UM	DBM0	10	atabase\imageFile\H	512,512	178,128	165,122	169,118	176,112	171,1
24	C2	CX40WT	NABU	500UM	DBM0	11	atabase\imageFile\H	512,512	22,140	49,122	37,106	62,104	62,9
25	C2	CX40WT	NABU	500UM	DBM0	12	atabase\imageFile\H	512,512	27,39	8,58	8,84	24,80	48,5

Fig. 2-4 ***Screen shot of experimental descriptor*** *PhaseConst_Exp_Setup.xls.*

The details of different attributes are self explanatory except for the 'Drug' and 'Drug Exposure' for which a unique nomenclature was introduced. In timelapse experiment sometimes it is given that a combination of drugs are given and again this combination is applied in a different regime. For example if drug X and Y were given as a mixture or simultaneously then they would be tagged as XaY, however if they were introduced sequentially then XfY, which translates after given drug X, drug Y was administered. Any number and combination of drugs can be tagged by this nomenclature. Dose of the drug is written in two capital letter words, e.g. NM means nano molar. If a combination of drugs were used corresponding doses will be written with 'a' or 'f' alphabet as used earlier.

Drug exposure reflects the duration (in min) the drug was present in the medium. In cytometric experiments drugs are administered in two ways - bolus and non-bolus (termed continuous). In a bolus scenario the medium is washed to remove the drug after a certain exposure time, while in non-bolus situation the drug remains in the medium throughout the experimental duration. Another major difference is the point in the experiment at which the drug(s) was administered i.e. before or after the start of the experiment. Based on these criteria of administration the following nomenclature was developed:

Table 2-2 drug exposure nomenclature (time duration is in minutes)

BBM (Bolus Before Mark)	BBM60t0	Bolus given 60 min before the start of the experiment.
BAM (Bolus After Mark)	BAM1440t1500	Bolus given 1440 min to 1500 min after the start of the experiment.
DBM (Drug Before Mark)	DBM120	Drug given 120 min before the start of the experiment.
DAM (Drug After Mark)	DAM1440	Drug given 1440 min after the start of the experiment.

Below is an example where two scenarios are encoded representing a unique drug, dose and exposure protocol.

Table 2-3 Exemplar drug dose combinations

Drug	**Dose**	**Exposure**	**Explanation**
A	1NM	BBM60t0	1 nano molar concentration of drug 'A' was given as a bolus 60 minutes before the start of the experiment. (for how long)
AaBfD	1NMa2NMf0.1NM	BBM120t0DAM1440	A mixture of drug A and B of concentration 1 nano molar and 2 nano

			molar was given as bolus for 120 minutes before the start of the experiment followed by drug D which was given at 0.1 nano molar concentration at 1440 minutes after the start of the experiment.

2.3.4 The encoding Folder

This folder contains the encoding tool - ProgeniTRAK. The software is divided into three interlinked parts. The first part interacts with the digital laboratory notebook (figure 2-4) and directs users to a specific progenitor cell location, this part of the software also generates the tag through which the progenitor cell becomes indexed. The second part interacts with MM and draws the evolving lineage on a canvas. Finally, the third part writes the image-derived parameters associated with each cell of the lineage into a tab-delimited text file. A brief description of the encoding process (for details see Appendix II) will be given later in this chapter.

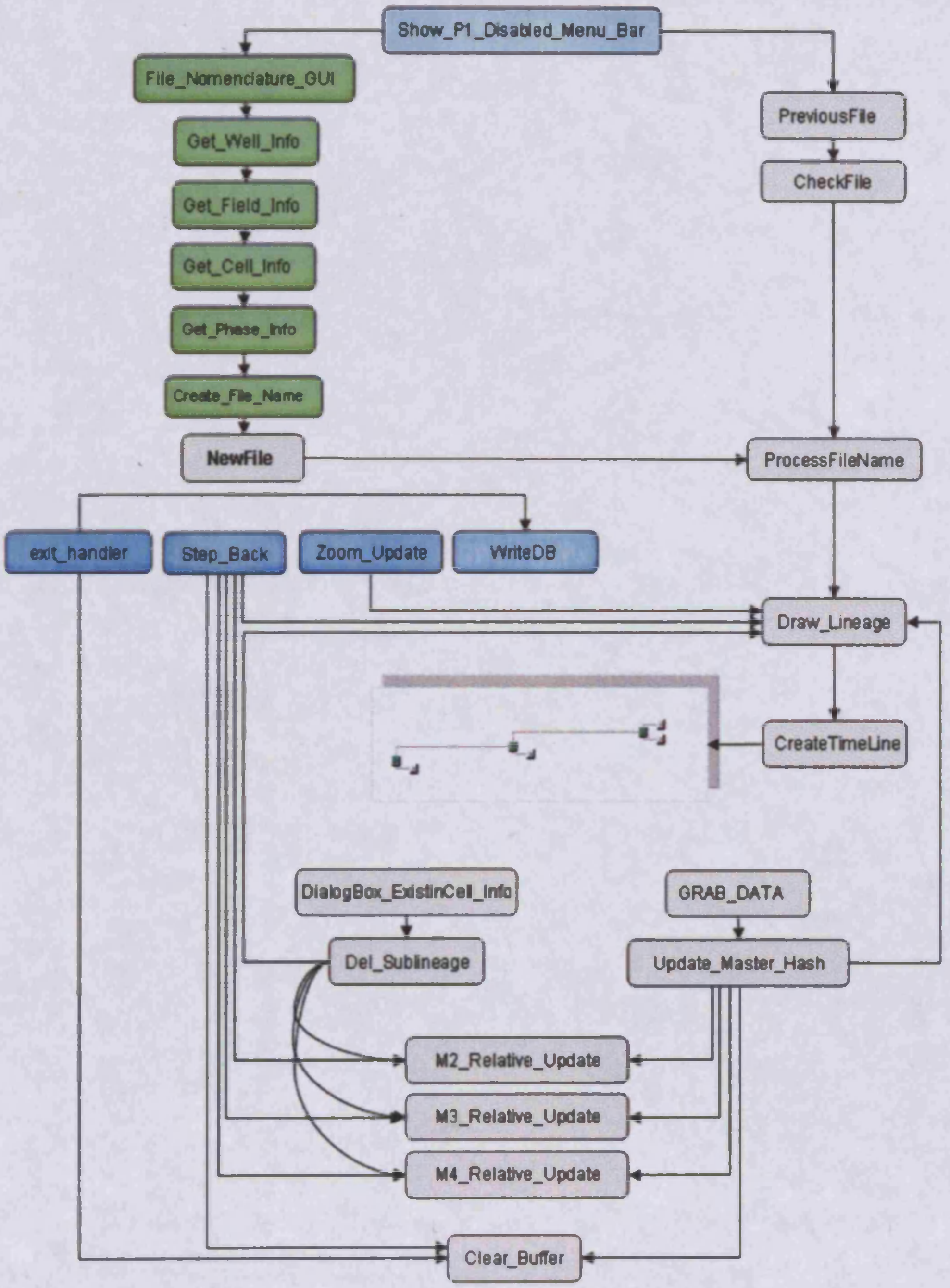


Fig. 2-5 ***Design and different functions of ProgeniTRAK*** *(details of the software along with the source code can be found in Appendix III).*

The aim of the encoding process was to encode the cell associated parameters from the image sequence in the specified data format, moreover during encoding, a graphical representation of the evolving lineage also needs to be displayed. The flow diagram below illustrates the lineage encoding process.

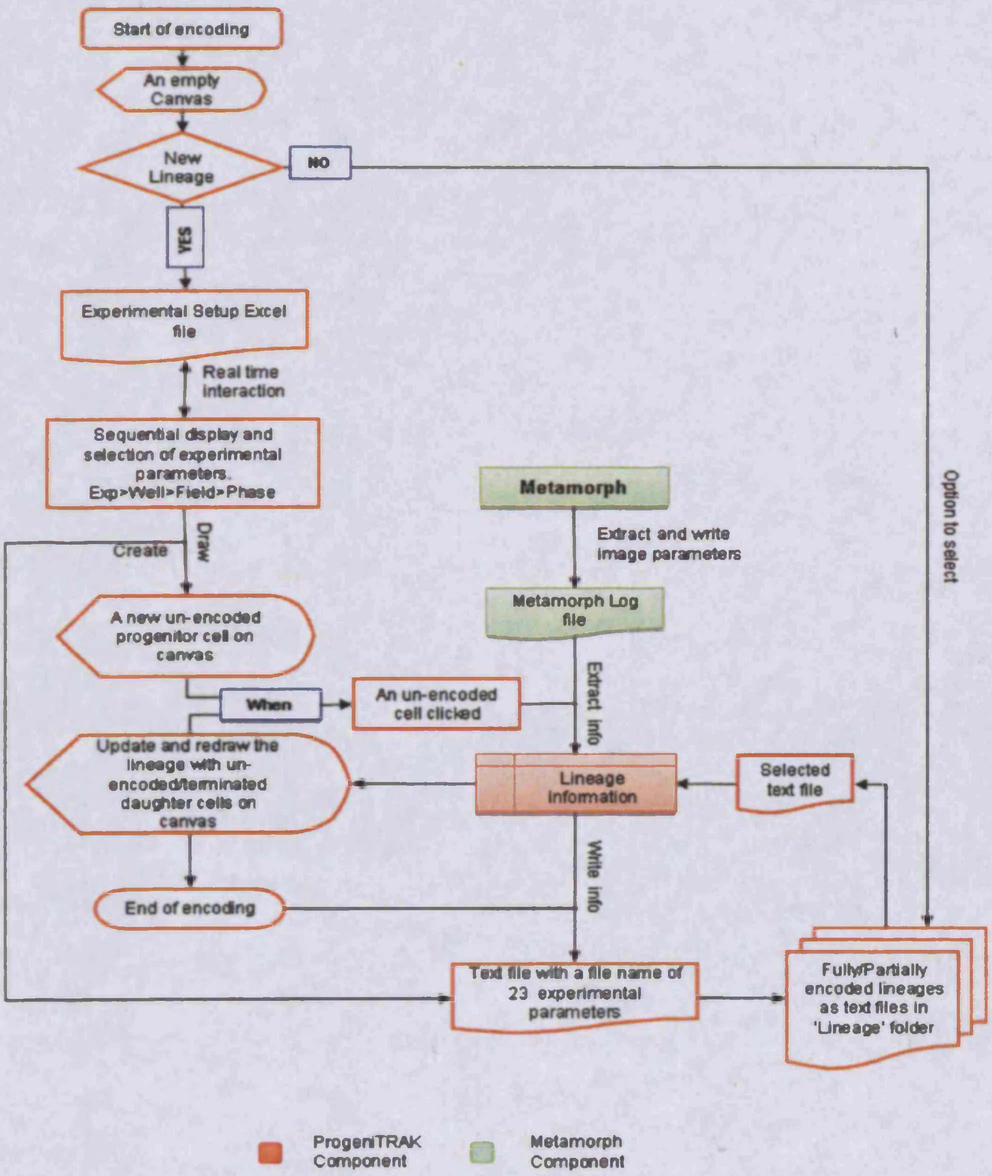

Start of encoding
An empty Canvas
New Lineage
NO
YES
Experimental Setup Excel file
Real time interaction
Sequential display and selection of experimental parameters. Exp>Well>Field>Phase
Create
Draw
A new un-encoded progenitor cell on canvas
When
An un-encoded cell clicked
Update and redraw the lineage with un-encoded/terminated daughter cells on canvas
End of encoding
Metamorph
Extract and write image parameters
Metamorph Log file
Extract info
Lineage information
Selected text file
Write info
Option to select
Text file with a file name of 23 experimental parameters
Fully/Partially encoded lineages as text files in 'Lineage' folder
🟥 ProgeniTRAK Component
🟩 Metamorph Component

Fig. 2-6 Flow diagram of the encoding process by ProgeniTRAK in conjunction with MM (highlighted as green)

Both MM and ProgeniTRAK are executed via a shortcut icon in the desktop (for detail of installing ProgeniTRAK see Appendix IV). At the start of the encoding process MM is opened by clicking the MM program icon on the desktop. In the journal folder six journal files have already been installed (see installation manual for detail in Appendix IV) with associated buttons in MM menu bar and will appear as follows.

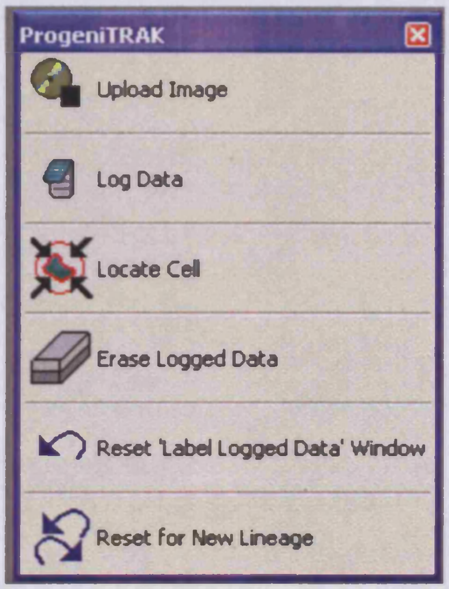


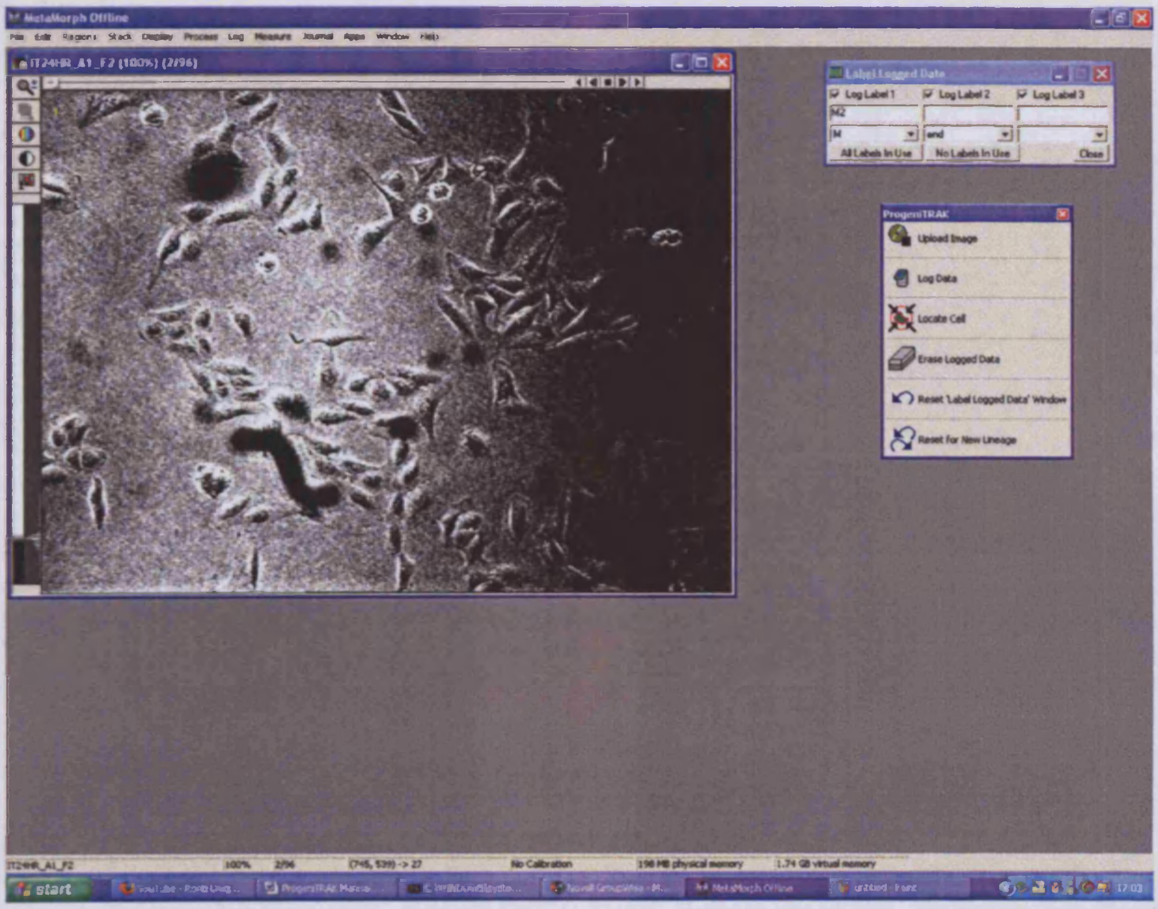
Fig. 2-7 ***Journal execution button at MM menu bar.***

Once these buttons were found, then user selects the 'Label Logged Data' option from the 'Log' menu of MM. In the 'Label Logged Data' window, the user checks the 'All Labels in Use' check button. In the 'Label Logged Data' window the first column should have the following tags in the pull down menu (second row) M2, M3,M4,P,R,D,L,S,V. If all tags are not available user types the tags sequentially. The tags represent all possible event that can occur to a cell, description of each tag is given in the following table:

Table 2-4 Event tags used in ProgeniTRAK

Tag	Description
M2	Mitosis 2 – Cell divides into two daughters cells
M3	Mitosis 3 – Cell divides into three daughters cells
M4	Mitosis 4 – Cell divides into four daughters cells
P	Polyploidy – Cell divides to single daughter cell
R	Refused – Two divided cell fused back to one cell
D	Dead – Apoptosis or Necrosis
L	Lost – When cell is lost from the FV
S	Survived – A living cell at the last frame of the image sequence
V	Unresolved – If the image sequences ends before the end of a ensuing event

In order to find the desired image sequence/file the “Upload Image” button (figure 2-7) at the menu option is clicked, which allows the user to browse the appropriate image file in the ‘Image Folder’. Once uploaded MM is set for lineage encoding and the window should look as follows:



☑ Log Label 1 ☑ Log Label 2 ☑ Log Label 3

M2

M end

IT24HR_A1_F2 100% 2/96 (745, 539) -> 27 No Calibration 198 MB physical memory 1.74 GB virtual memory

Fig. 2-8 ***Screen shot of MM as it is set for encoding.***

2.4 Creating a new progenitor cell

During the encoding process, MM provided all the facilities for image viewing, data extraction and parsing as described earlier. The encoding and organizing of the lineage were performed by ProgeniTRAK and was executed once MM was set for encoding. ProgeniTRAK was executed by clicking the icon provided at the desktop. In order to setup ProgeniTRAK for a new lineage, six sequential steps were followed that defines the process of encoding:

Step 1: Users select the 'New' option from the menu of ProgeniTRAK, this invokes a second GUI with the option for the selection of experiments. ProgeniTRAK interacts with 'PhaseConst_Exp_Setup.xls' in real time and displays all the experiments registered in the Excel file.

Step 2: From the drop down menu users select the appropriate experiment, the image file opened in MM should belong to the same experiment. Once selected this experiment window disappears and a new Well window appears with all the Well information for the selected experiment. Again ProgeniTRAK reads the information from the Excel file.

Step 3: The Well window is a graphical representation of multi-well plate, where each square represents a well and the associated information (cell type, drug etc.) is written on each Well. Also in the header the experiment level information are displayed and the title of the window displays the experiment name. Users select the Well within which lies the field of view, this in turn invokes the disappearance of the Well window and appearance of field of view (FV) window.

Step 4: The FV window gives the option of field numbers that belongs to the selected well and users select the appropriate field and this in turn invokes a digital representation of the FV.

Step 5: The digital FV contains small squares colour-coded buttons representing each progenitor cell. Violet square represents a completed encoded lineage, orange - a partially encoded lineage and grey represents an as yet uncoded lineage. It is important that this digital representation of progenitor cells resembles with the first frame of the actual image file opened at MM and once assured users select the desired progenitor cell to be encoded as a full lineage map by clicking on the cell.

Step 6: This final window gives users the option of attributing the cell cycle phase of the progenitor cell to be encoded. By default it is set as unknown 'U', but in certain experiments this can be better defined. When the selection is complete, a tag or header is assigned to the selected progenitor cell and a digital representation of the progenitor cell (in grey colour) is created in the canvas of ProgeniTRAK. The tag or header of each progenitor cell has 23 parameters associated with it, which makes it unique against all other progenitor cells encoded via ProgeniTRAK.

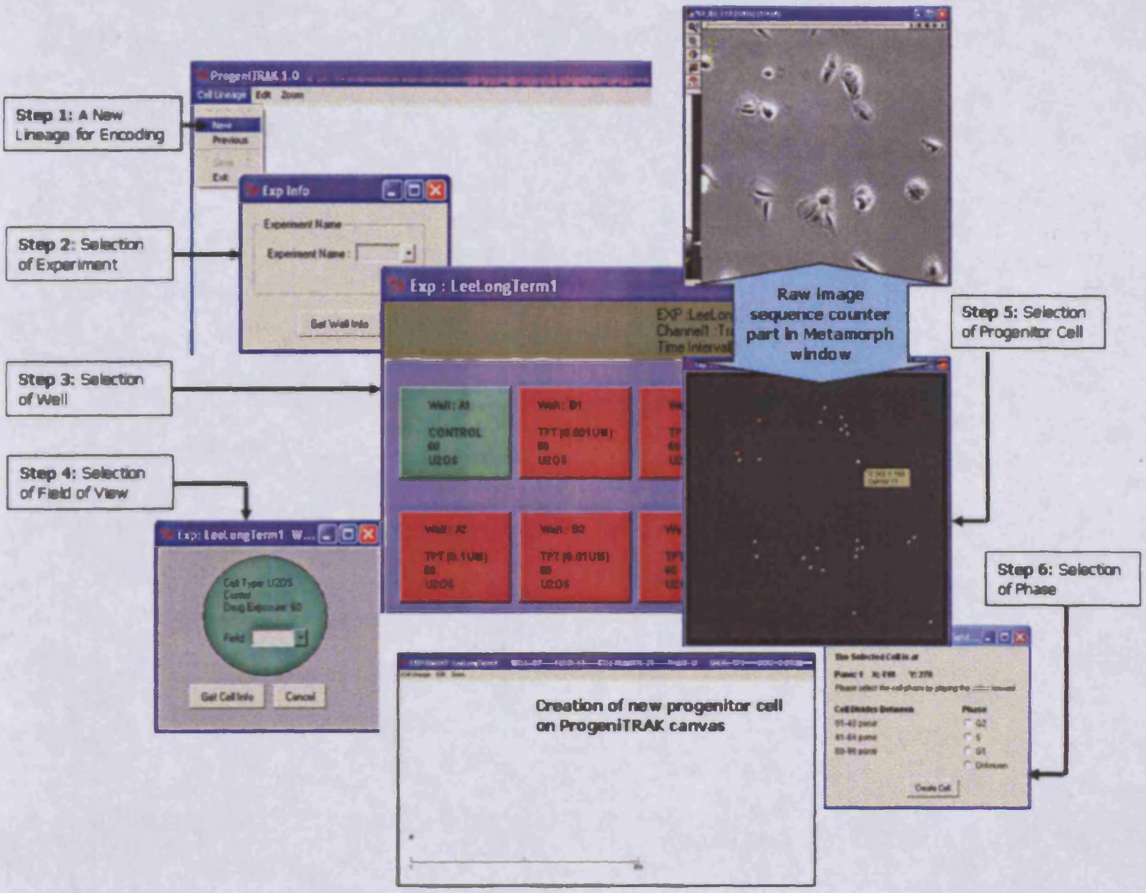
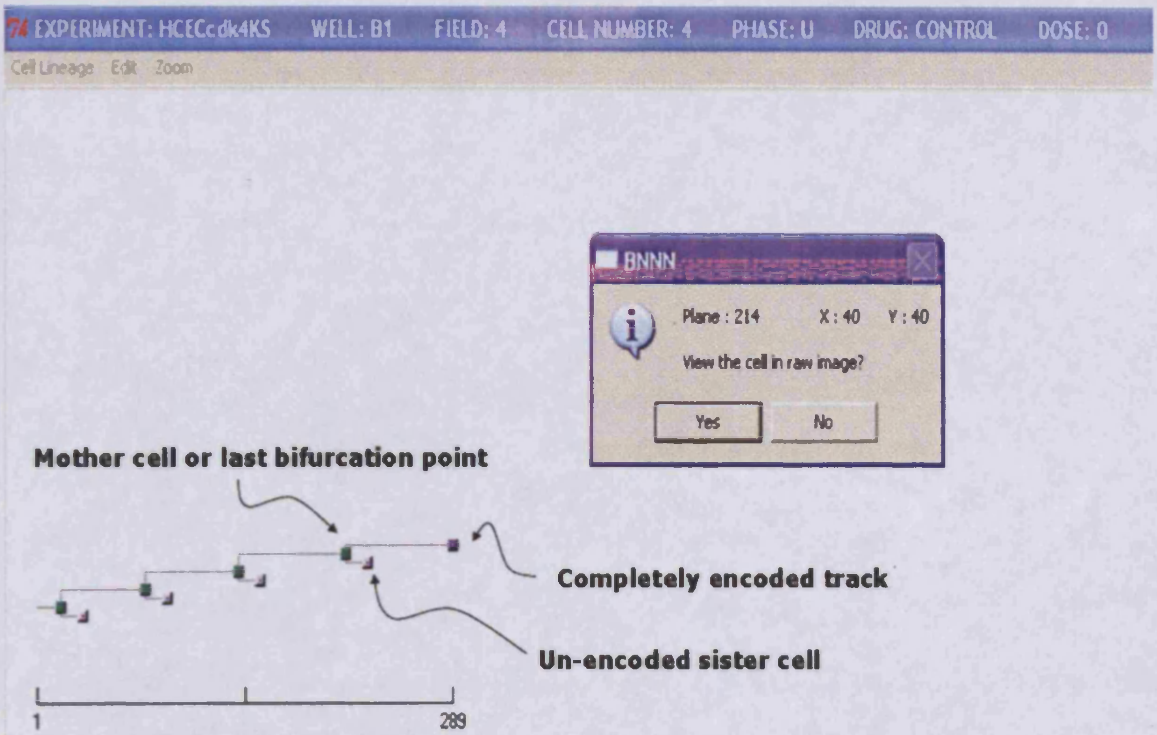


Fig. 2-9 ***The principal six stages of encoding.*** *Illustrated above are represented with their associated GUI than leads towards generating a new cell lineages in the ProgeniTRAK canvas.*

Since the header or the progenitor cell tag does not involve any manual typing, it is assumed to be less error prone, more over this sequential GUI representation orientates users in the experiment scenario including all the pertinent details and makes the data sharable amongst users. The nomenclature format means that anyone can interpret and negotiate the experiment, in other words it is not user specific. The progenitor cell tag was then used to name the text file providing the details of each encoded lineage. Constructing the file name or progenitor cell header through this sequential selection process and via a GUI structure ensures consistency of nomenclature and data encoding quality but also helped users to orient within the experimental context. This approach is considered to have a generic use and can be adopted for other experiments which does not even have to be image based since this approach mimics the 'laboratory note book' theme where all experimental descriptors are written and visited later for extracting or selecting certain information.

As described earlier a lineage comprising of one or more cells evolves from a common progenitor cell. Once the progenitor cell is created in the canvas of ProgeniTRAK Window (PTW) with the appropriate header tag, users start to follow/track the actual real counter part (actual cell) in the image (MM window; MMW) by playing the image sequence. When an event occurs (see figure 2-10) to the cell under tracking, the playing of image sequence is stopped and then rewend to the start point of the event which is usually a few preceding frames. To exemplify, if the cell under analysis goes through mitosis, the start of the mitosis is identified as the point of cell rounding. In the 'Labeled Log Data' window of MMW the tag 'M2' is selected and in the image a ROI is placed on the nucleus of the cell and then the button "Log data" (figure 2-7) is clicked. The image sequence is forwarded up to the point when the cell splits into two daughter cells and again at the middle point the ROI is placed again and the button is clicked again. With each click, the ROI extracts 24 parameters from the image, and these include - Image Name, Image Plane, Image Date and Time, Elapsed Time, Stage Label, Wavelength, Z Position, Region Label, Area, Distance, Angle, Left, Top, Width, Height, Threshold Area, Threshold Area %, Threshold Distance, Average Intensity, Intensity Standard Dev, Intensity Signal/Noise, Integrated Intensity, Minimum Intensity and Maximum Intensity. The time, position and intensity parameters are pivotal for a wide range of cell based assays. In addition to the aforementioned parameter pairs, 5 further parameters are also encoded for each event and include cell name, event type, step number and canvas coordinates, 5 blank spaces are also encoded for each cell so that future annotations can be incorporated. Altogether these 24x2+5+5 = 58 parameters constitutes a row of data fields within the text file. Once these parameters are parsed to ProgeniTRAK via DDE link, ProgeniTRAK can acknowledge the event type and accordingly redraw the lineage, for this mitosis example two new cells appear in the canvas with two connecting lines. The north daughter is followed in the same manner and once a further event occurs the ROI is placed and parameters get encoded, the process continues until the end of a track is reached, i.e. tracking till the end of the image sequence.



EXPERIMENT: HCECcdk4KS WELL: B1 FIELD: 4 CELL NUMBER: 4 PHASE: U DRUG: CONTROL DOSE: 0

BNNN

Plane : 214 X : 40 Y : 40

View the cell in raw image?

Fig. 2-10 Visualizing and encoding of an evolving cell lineage in ProgeniTRAK canvas.

Once a track is fully encoded, in order to encode the sister track, the mother cell of the last cell in the track is clicked which displays the frame number and the coordinate of the mother cell. By clicking the "Locate cell" button in MMW (figure 2-7), the user inputs the information which then indexes the mother cell in the image sequence and from that point the sister cell is tracked and encoded in the same manner. Once a lineage is complete (all tracks ended) then the save option in PTW ensures the whole lineage is saved as a text file in the 'Lineage Folder'.

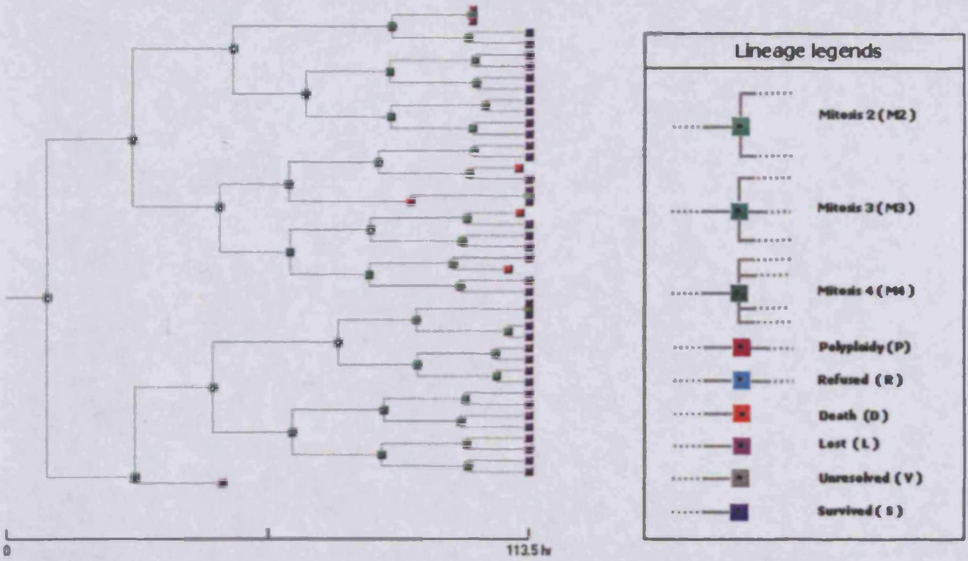


Fig. 2-11 A complete lineage encoded from a real progenitor cell spanning up to five generations. Associated legends describe possible outcomes and tags which describe the cellular outcome.

Setting the ROI and selecting the type of event manually is indeed a rate-limiting step, but the user interaction ensures the highest quality of data encoding particularly the outcome of cell mitosis. A combination of automated and user-interactive bioinformatics software is what is suggested in a recent review (Giuliano et al. 2005) as the challenge and opportunity for next generation of high content screening (see chapter 7 for further discussions). At this point it is important that the quality of the encoding is high and this was best ensured by manual tracking. In order to encode another lineage from the same field of view, users does not need to go through the selection process again but rather just click “Reset button” (figure 2-7) in MMW, which invokes the digital FV again and users start the process from that point onwards. If an error is made and the user realizes the mistake the encoding can be corrected by using an “Erase button” in PTW which removes the last two entries in the log file. However if the user realizes the mistake earlier then the “Step Back” button in the PTW will redo the event encoding again. Through ProgeniTRAK it is also possible to encode and save a lineage partially and finish the encoding process at a later time; it is also possible to delete any portion of the lineage and edit the information as required.

2.4.1 The lineage Folder

As of January 2008, 745 lineages were completely encoded from three separate experiments. The lineages were encoded by a wide user group and include - Drs. Lee Campbell, Nuria Marquez, Janet Fisher and Marie Wiltshire. Each lineage was saved as tab delimited text file, where the file name has 23 underscore separated parameters facilities to index the lineage or progenitor cell while in the text file each line comprise of 58 data fields separated by tabs. All three experiments represented drug screens where anti-cancer agents were added to human osteosarcoma cells (U-2 OS cells) (for detail see Appendix V). At different time intervals the lineages were subjected to a visual and automated quality control check. This was done to ensure that while encoding users mistakenly did not assign the wrong tag to the cellular event. Small PERL scripts were written to ensure the integrity of extracted parameters while random visual check on events ensured the rigor and robustness of the manual encoding and assignment of the event tag. The lineage folder is considered as the primary database (called as 'LDB' hereafter) upon which preliminary data analysis was performed, this folder also facilitates various hypothesis driven data mining and provided a shared access to this folder, it was placed onto a shared drive on the network. Importantly the users were not able to place the data directly into this database - this was handled by a designated database master who would check the data quality and then enter the new data into the database.

2.5 The LDB inventory

Below is a summary of the database inventory, the first two rows represent the lineage data derived from two dose dependent topoisomerase I inhibitor screens (Feeney et al. 2003) (i.e. topotecan) (Kollmannsberger et al. 1999), while the bottom row represents a different topoisomerase inhibitor screen (topoisomerase II) with ICRF and Taxol® screen.

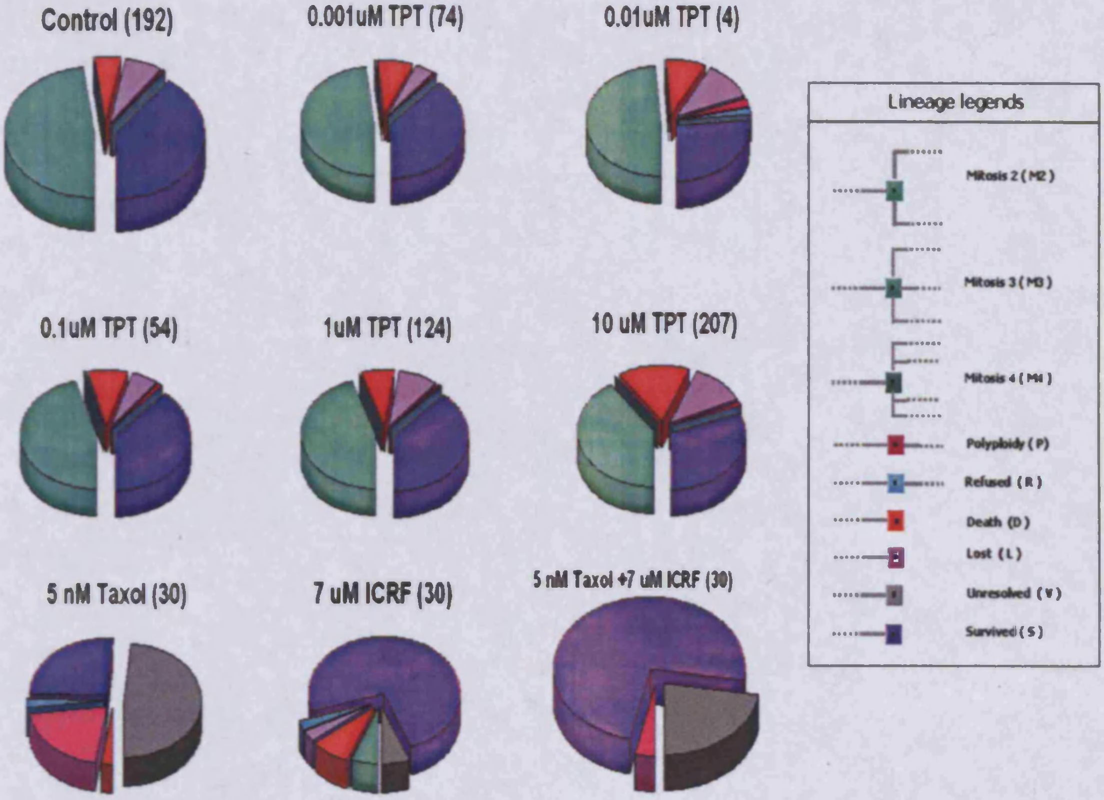


Fig. 2-12 ***The LDB inventory****. Each pie chart represents a drug and dose as stated in the title of the pie chart and in the parenthesis the number of lineages encoded for that particular dose is mentioned. The purpose of the pie charts is to give an overall review of the predominant events associated with each treatment regimen. The colour represents individual events (event legends on right) and computed as percentages of the total events encoded for each dose.*

Further analysis tabulating the events illustrate the actual number of each event against the number of lineages encoded was performed.

Table 2-5 Number of lineages and events at each dose.

	Control	0.001 uM TPT	0.01 uM TPT	0.1 uM TPT	1 uM TPT	10 uM TPT	5nm Taxol	7 uM ICRF	5nm Taxol+ 7uM ICRF
Lineage Num	***192***	***74***	***4***	***54***	***124***	***207***	***30***	***30***	***30***
Mitosis 2	2822	1531	22	366	625	457	0	2	0
Mitosis 3	8	6	0	4	1	1	0	0	0
Mitosis 4	1	0	0	0	0	0	0	0	0
Death	274	247	4	69	98	177	1	3	0
Lost	417	139	5	47	122	132		1	0
Polyploid	16	11	1	5	8	12	8	0	1
Refused	6	2	1	1	4		1	1	
Survived	2335	1224	12	309	531	350	10	26	23
Unresolve	6	0	0	0	0	0	19	2	7

From the inventory alone a dose dependent event pattern becomes evident, for the first two rows the percentage of cell death events increased with dose and furthermore upon Taxol® treatment the number polyploidy event increased.

Before setting out to undertake comprehensive or high-level data mining and analysis it is pivotal to validate the encoded data by performing simple data mining on the encoded data and to compare the database output results with results that were generated from completely different processes. Counting different cellular events manually is one of the common practice in cell based assays (Feeney et al. 2003; Marquez et al. 2003). The objective of such an endeavour is to establish the effect of drug and dose on the generation and/or abrogation of cellular events. Through this simple inventory, it was possible to quantify and visualize the drug and dose dependency on cellular event variation. However a manual vs. encoded cellular behaviour comparison was required to validate the data format before further mining and analysis; to this end Cell Density (CD) was identified as a suitable time-dependent output parameter because a CD index provides a measurement to quantify the level of artefact or noise dependent on cell movement particularly when counting cells manually. In timelapse experiments, living cells move due to their innate behaviour and this often causes migratory cells that were initially outside the field of view to move within the field of view. For instance a highly migratory cell moves into a field of view undergoes a mitotic event and these events or resultant cells are counted contributing noise to a straight cell count, therefore the assay is unbounded and is highly variable (this is particularly effected when anti-cancer agents have a dual effect on cell migration and cell proliferation i.e. the effect of Taxol®). However when using ProgeniTRAK, cells are not included in the assay if they appear mid-time course and as such noise is not added to the count. Validation was performed by comparing the cell density index derived from the lineage database and that obtained from manual counting. At a simple level the greater the difference between the cell density derived from the manual count and those derived from the lineage database, the higher the motility of the cells would be. Primarily the ability to compare provided the validity of the data format and secondly through the cell density index the extent of noise could be verified. Six FVs were selected from the two replicate screens using topotecan where all FV had an area of 512x512 pixels (approximately 400x400 µm) and contained U-2 OS cells in control conditions. The population count within this area at a particular time point was the basis of the CD index. CD indices for six FVs were measured at 113 h both manually and from encoded lineages and the difference is plotted in the following graph. CD index for each FV was measured by:

$$CD\,index_{\,t} = \frac{N_{\,t}}{512 \times 512} \times 100 \qquad \text{Eq. 1}$$

Where N is the number of cells present at time t=113 h of the experiment.

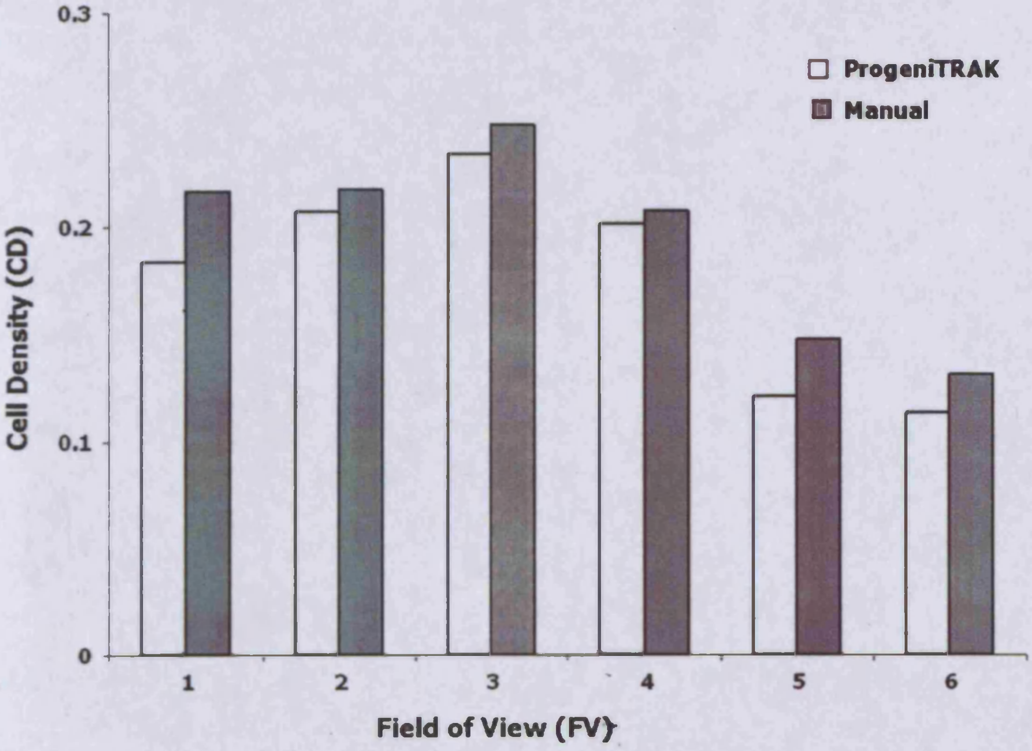


Fig. 2-13 ***Cell density measurement.*** *Bar plot showing cell density difference measured from encoded lineages (white) and by manual counting (grey) at the 113 h in six different FVs.*

From the figure above, it was evident that in all instances the manual measured density was always an over estimation (with an average of 0.018 cells/100 pixel^2) and can be attributed to the counting of migratory cells, i.e. noise. Although it is possible to ignore the contributing it would be difficult to compare cell types, and indeed the perturbing agent may affect the motility of cells. A series of data mining and analysis were performed which will be presented in the subsequent chapters which not only represented some common analytical procedures and results but also highlighted routes for revealing otherwise occult cellular behaviour. The knowledge gained from these analytical results when compared with the priori knowledge regarding the cell cycle action of topotecan (Feeney et al. 2003), showed a high degree of resemblance and this aspect was part of the validation of the data format as well as the infrastructure itself.

2.6 Concluding remarks - data mining using the lineage format

In simple terms the lineage can be viewed as a bifurcating (in most cases) map where each node represents an event while the line connecting two consecutive nodes represents time duration for a cell cycle and is termed as inter mitotic time (IMT) as it is the time between two sequential mitotic events (i.e. between mother and daughter cells). As a cell usually divides into two daughter cells, the mitosis of the mother cell could be colloquially referred to as the birth of the daughter cell, so that the IMT represents a complete cell cycle and would encompass all the phases of the cell cycle. For ProgeniTRAK no information was encoded according to the cell cycle time per se i.e. along the node connecting line and the only data being encoded was when an event occurred i.e. node-to-node. Each node also contains the start and end information of the event. It is important to note that IMT values were not encoded rather these are calculated on-the-fly during data mining routines and require two successive mitosis. If however the daughter cell undergoes cell death then the line connecting the mother and daughter cell does not represent IMT as the daughter cell did not undergone a complete cell cycle, the same holds true for any other event (lost, survived and unresolved); except mitosis (M2, M3 and M4).

Lineages provide rich multi-dimensional data, where the preliminary level involves vertical segmentation of the lineage and another level involves horizontal segmentation. Vertically each linage can be segmented in relation to time or generation along the x-axis and correspondingly in the y-axis the distribution of events can be measured.

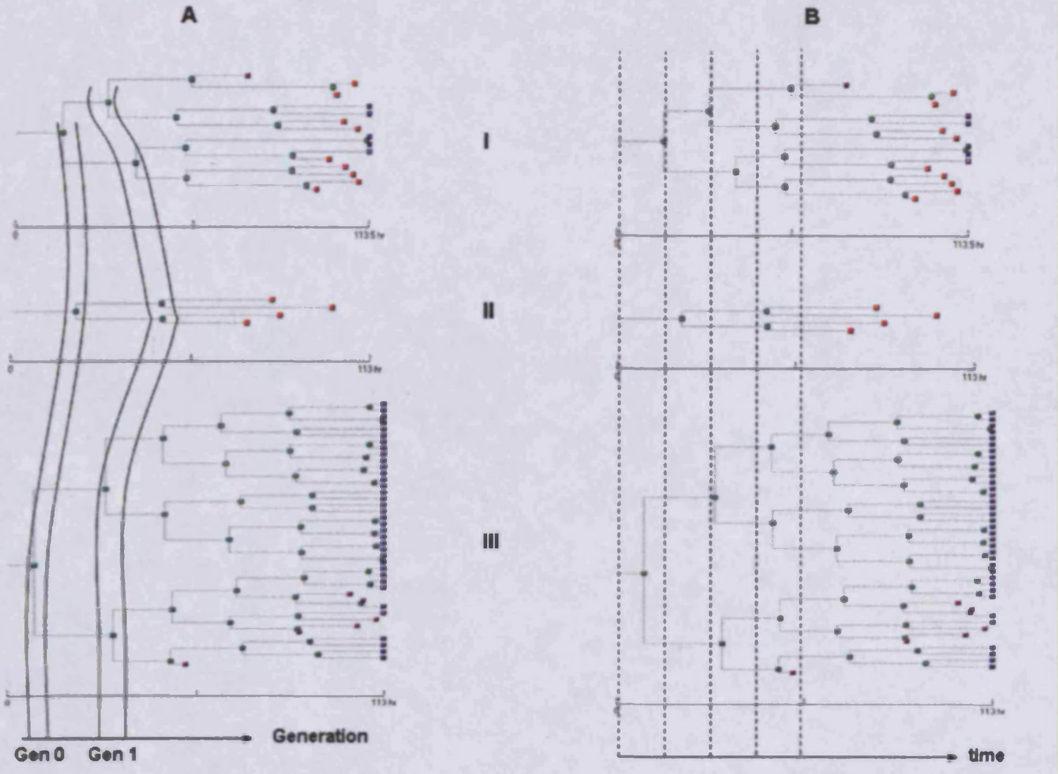


Fig. 2-14 ***Vertical segmentation of lineages****. Three exemplar lineages (I, II, III) were segmented vertically, A – in relation to generations and B in relation to time. During data mining usually a set of lineages are filtered according to a set of criteria, i.e. drug and dose. For each lineage the type and number of events can be counted in relation to the generation (solid line) or time (dotted line) and as the number of different events accumulates, it can be presented as cumulative curve or distributions. It is not only the event count but also other measured values like the inter mitotic time that can also be presented as a distribution (discussed later).*

Lineages can also be segmented horizontally termed as tracks (i.e. direct lines of descent), which provide temporal resolution for cell division and cell death timing, forming a bifurcation map with nodes, branches and cul-de-sacs. Track-dependent information depicts multi-cycle dynamics of cellular behaviour in a relationship context. Variability of IMT when analysed in such a relationship context could provide a dose and drug class-dependent pedigree behaviour. Moreover it is not only the variability of IMT but also the displacement that can be quantified and utilized to illustrate cellular behaviour. Thus, each node of the lineage not only encapsulates time and event but also the coordinates, which can be utilized to measure *inter mitotic displacement* (IMD) which represents the total displacement (in pixels or microns) during one complete cell cycle (an event-to-event measurement). Like IMT, IMD also is not directly encoded but can be calculated in real-time from the encoded lineages.

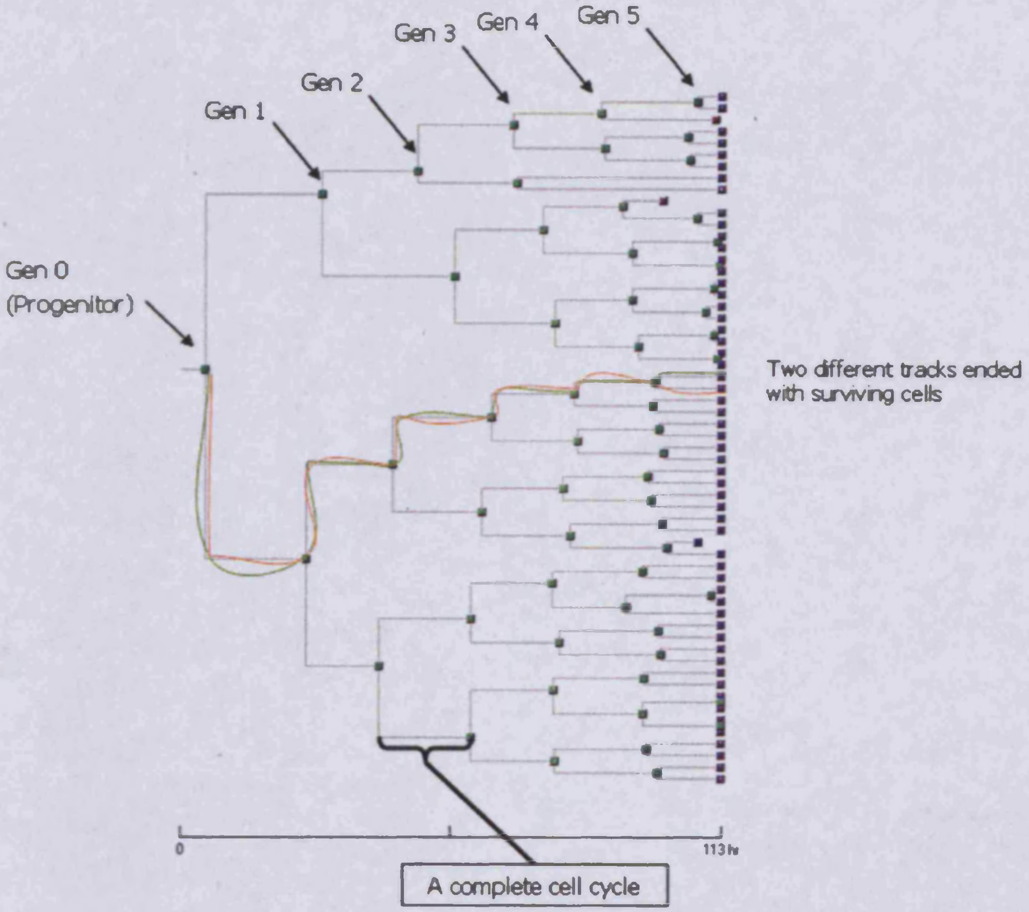


Fig. 2-15 ***Horizontal segmentation of a real lineage****. Arrow showing a single track through different generations where each node represents an event. Hand drawn lines (green and orange) depicts two tracks of the lineage. IMT is represented as 'A complete cell cycle' and displacement of cell during this time period is termed IMD. According to generation IMT and IMD is designated as IMTn and IMDn where n is the generation number which is always >0, since progenitor cells (generation 0) do not have a complete cell cycle duration encoded i.e. the experiment started after the birth of the progenitor cell.*

The relationship based information that a lineage provides when segmented horizontally is indeed a unique perspective to understand cellular dynamics temporally as well as spatially. The successive chapters explore this rich data source with an increasing complexity. The LDB is the primary source for these types of analyses and the initial analysis involves vertical segmentation of the lineages (see Chapter 3). Since event based analysis are predominant analysis techniques in cell based assays, where events are counted manually (as it happens in image sequence) and plotted in relation to time. These analyses simultaneously offered validation as well as demonstration of the pertinence and usefulness of such data format to elucidate cellular behaviour. Further advanced analysis based on horizontal segmentation of lineages was undertaken to elucidate patterns of inter-nodal track dimensions and lineage asymmetry, providing insights for drug resistance (see Chapter 4).

Chapter 3: ProgeniDB – A web-accessible lineage database

This part of the work was carried out in collaboration with Peter Husemann, a visiting scholar from Bielefeld University, Germany.

3.1 Introduction

ProgeniTRAK has provided a systematic framework for converting images to numbers producing cell lineages objectively parameterised with unique tags. These encoded data were archived in tab delimited text files within a folder structure - LDB, however this simple format limits the accessibility of the data repository to local users only. Initially, the text files were stored on a shared drive, accessible by a small number of individuals at Cardiff and these text format did provide the means for validating the approach. The current chapter focuses on the next phase of development, to produce a web accessible database (ProgeniDB) with associated data mining tools enabling a public accessible database. Therefore the premise for building such a database was to give public access to the novel lineage data and data mining tools to segment cellular event data according to imposed experimental and biological constraints. This aspect of the work involved design and development of the database along with user interface for selecting and filtering the data to download. Furthermore the downloaded data were analyzed to interpret cellular behaviour. Simple validation studies were undertaken to ensure that database retrieval led to event analysis outputs that matched the original analysis approaches. Studies were performed to obtain the simple kinetic behaviour of the population.

3.2 Design and implementation of ProgeniDB

Peter Husemann, the visiting scholar, brought expertise in web enabled database development utilizing cgi, PERL and MySQL Using the existing ProgeniTRAK derived parameters an ER diagram was designed to imitate the text format hierarchy explained in earlier chapter and facilitated the development of the relational database – ProgeniDB (Khan et al. 2007), that utilized the MySQL (see http://www.mysql.com/) database management system. This database had 5 separate tables defined as Experiment, Well, Lineage, Cell and Relations. Except for the RELATION table which basically stored the primary key for each table and LINEAGE table which represent the progenitor cell, all others tables were a representation of the data format introduced in figure 2-2.

The attributes for each table were illustrated in the following ER diagram where PK refers to Primary Key and FK, Foreign Key. Each cell that participated (including the progenitor cell) was given a unique ID and the ID number was given in incremental order, the same applied to Experiment, Well, Lineage IDs.

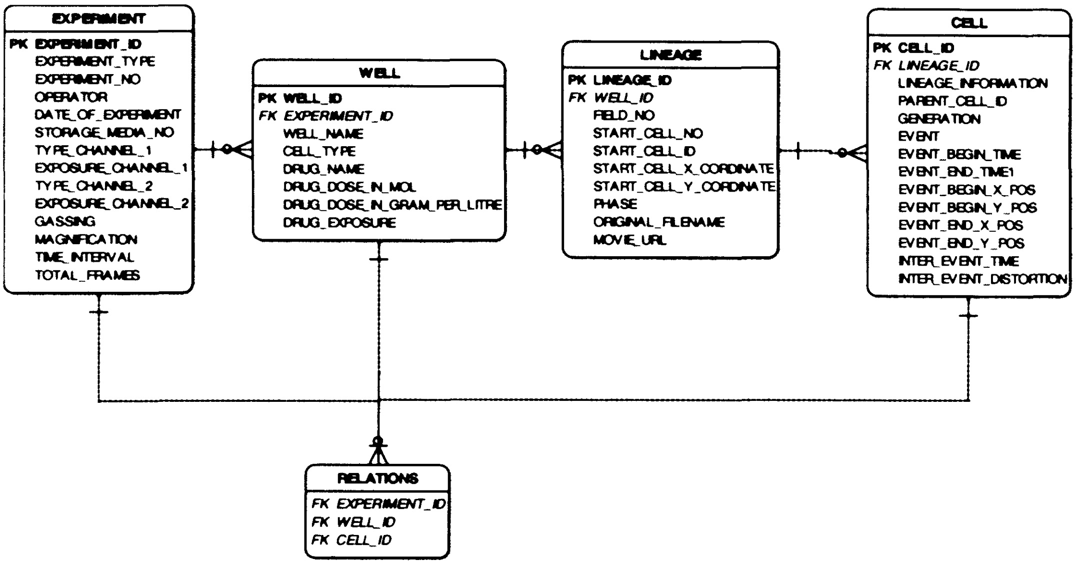


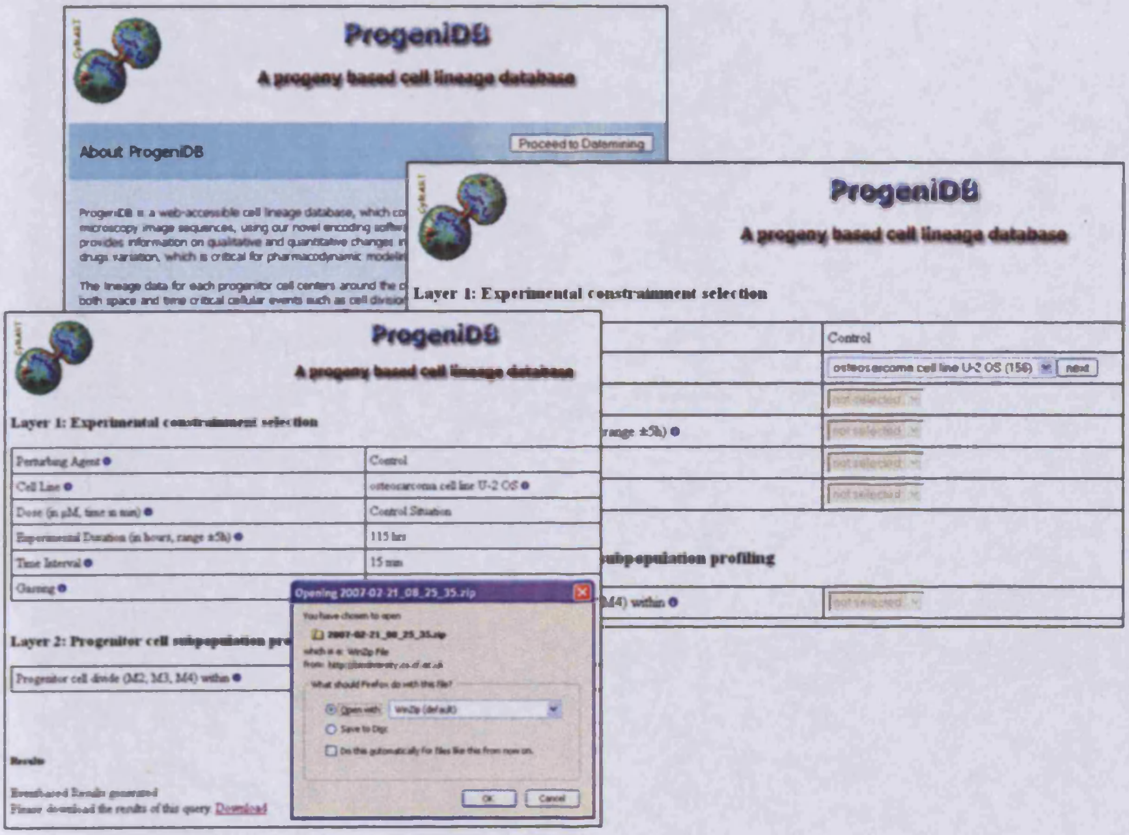
Fig. 3-1 ***ER diagram of ProgeniDB.***

A MySQL script called *database_tables.sql* was created which when first executed goes through a process of deleting the whole database and then re-creating it according to the ER diagram. Once created another PERL script *populate_database.pl* was executed to populate the database with data extracted from the text files held in a Lineage folder. This process guarantees that with every quarterly update the cell IDs do not become redundant as the Lineage folder holds previous as well as newly encoded lineages as text files.

At present ProgeniDB has over 622 lineages with 12,560 cells it is accessible through any standard web browser where a logical query and download of data can be achieved. The encoded lineage data at LDB were passed through rigid quality control as described in earlier chapter as such ProgeniDB ensured quality control passed lineage data. At present use of ProgeniTRAK was limited to a small number of users but a future growth was anticipated that invokes rigid protocol for encoding and centralized quality control, where lineages will be encoded at multiple sites following a rigid protocol and deposited to a central place where quality control check will be performed before archiving into ProgeniDB.

3.3 Mining and analyzing ProgeniDB

Generally, data mining (sometimes called data or knowledge discovery) is the process of interrogating data from different perspectives and summarizing it into a format for interpretation. Data mining is an analytical approach which enables users to analyze data from many different directions, and represents a multi-dimensional space that can contend with larger data sets, with less 'danger' for making inferences, and often uses data that was originally acquired for a different purpose or from a different perspective. A data mining approach differs from data analysis which is the process of looking at and summarizing data with the intent to extract useful information and develop conclusion (Abbott et al. 1998; Thearling 2008). Here both mining and analyzing ProgeniDB data were demonstrated, where the overall objective of the ProgeniDB mining process was to select (filter) a set of lineages based on a defined experimental or biological constraint with a dynamic web-page to guide users to query the database in three progressive stages. The first page of ProgeniDB gave user information including an overview of phase contrast microscopy and a diagrammatic outline of the lineage encoding process along with nomenclature and terminology used for defining a lineage structure. Informed users may then select to proceed to querying of the database by clicking the 'Proceed Data Mining' button. This button leads to the query page where the first stage of the query process is selection of experimental conditions, which includes the perturbing agent (typically a drug), the cell line, drug dose, experimental duration, sampling interval and plate gassing (typically 5% CO_2 or N_2). These experimental filters can only be selected one at a time and in a sequential order. Importantly, options within a current filter will depend on all previously selected filters, for example options of 'cell line' will depend on the previous selection of perturbing agent, therefore the user cannot select a combination of filters that don't exist in the data. An information icon was provided alongside each filter which showed an explicit description of that filter if the user hovered above the icon. Also provided within brackets is the number of lineages selected under the filter conditions. These 6 consecutive selections led to the second stage of the query process – called progenitor cell sub-population profiling. By default users were able to skip this stage of selection by pressing 'next' button, however this stage has two implications on the query process. First, from the drop down menu, it was possible for users to obtain the number of progenitor cells that deliver to mitosis against time. This effective count leading to rate of mitosis was an important output for obtaining cell cycle-related measurements of drug effects on progenitor cells as exemplified in the case study. The second implication was that through such profiling, it was possible to select a sub-set of lineages based on progenitor cell cycle position (or age) and to quantify the downstream impact of drug treatment at that cell cycle age on successive generations.



ProgeniDB

A progeny based cell lineage database

About ProgeniDB

Proceed to Datamining

ProgeniDB is a web-accessible cell lineage database, which co microscopy image sequences, using our novel encoding softw provides information on qualitative and quantitative changes in drugs variation, which is critical for pharmacodynamic modelin

The lineage data for each progenitor cell centers around the c both space and time critical cellular events such as cell divisio

ProgeniDB

A progeny based cell lineage database

Layer 1: Experimental constrainment selection

	Control
	osteosarcoma cell line U-2 OS (156) next
	not selected
range ±5h) ●	not selected
	not selected
	not selected

subpopulation profiling

M4) within ●	not selected

ProgeniDB

A progeny based cell lineage database

Layer 1: Experimental constrainment selection

Perturbing Agent ●	Control
Cell Line ●	osteosarcoma cell line U-2 OS ●
Dose (in μM, time in min) ●	Control Situation
Experimental Duration (in hours, range ±5h) ●	115 hrs
Time Interval ●	15 min
Gating ●	

Layer 2: Progenitor cell subpopulation pro

Progenitor cell divide (M2, M3, M4) within ●

Results

Event-based Results generated
Please download the results of this query Download

Opening 2007-02-21_08_25_35.zip

You have chosen to open

2007-02-21_08_25_35.zip

which is a: WinZip File

What should Firefox do with this file?

◉ Open with: WinZip (default)

○ Save to Disk

□ Do this automatically for files like this from now on.

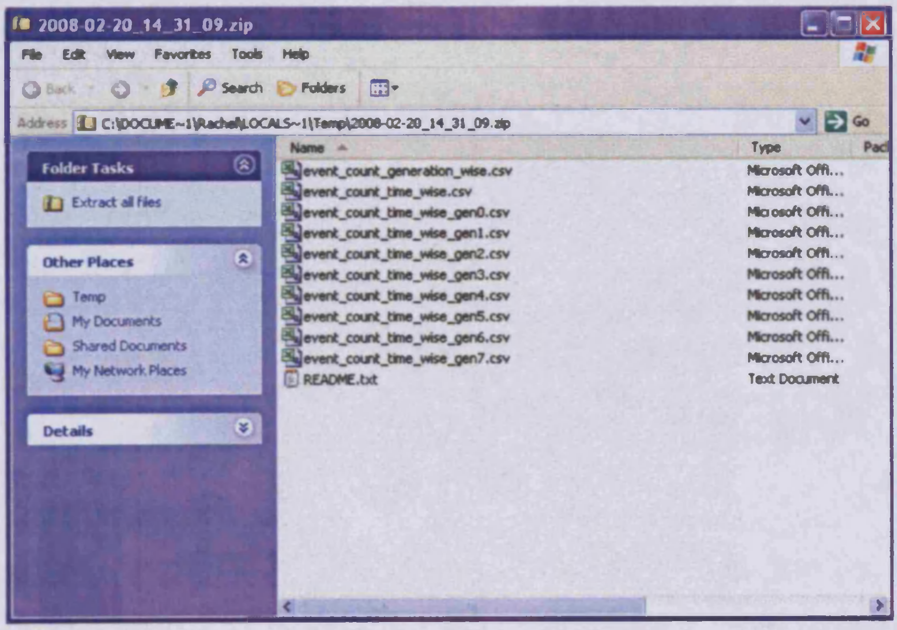
OK Cancel

Fig. 3-2 Screen shots to illustrate both the user interface and the logical query process through ProgeniDB.

Based on the filter selections at the previous two stages, the final query page displayed a list of experiment ID(s) where each experiment included the lineages that satisfy the selected filters. Important experimental attributes were displayed in a table format which gave a preview about the experiment(s), however if users wanted a detail description about the experiment, an information icon was provided along side each experiment which when clicked would give a detailed description about that experiment. By default all experiments were selected, but users can choose any number of the experiment(s) from the list. At the end of the query stage when the 'Generate Results' button was pressed, 10 CSV (comma separated values) files are generated along with a README file. These files can then be downloaded as a zip-archived bundle by following the given download link.

3.3.1 Downloaded results

The downloaded zip-archived bundle contained event distribution data in relation to time and generation or in combination. As illustrated in the previous chapter lineages can be mined or segmented vertically as well as horizontally. Vertical segmentation can generate data that depict distributions of various nature with event distribution being predominant. At present in ProgeniDB only event distribution can be mined and the resulting CSV files can be subjected to data analysis for understanding cellular behaviour.



2008-02-20_14_31_09.zip

Address C:\DOCUME~1\Rachel\LOCALS~1\Temp\2008-02-20_14_31_09.zip

Name	Type	Pac
event_count_generation_wise.csv	Microsoft Offi...	
event_count_time_wise.csv	Microsoft Offi...	
event_count_time_wise_gen0.csv	Microsoft Offi...	
event_count_time_wise_gen1.csv	Microsoft Offi...	
event_count_time_wise_gen2.csv	Microsoft Offi...	
event_count_time_wise_gen3.csv	Microsoft Offi...	
event_count_time_wise_gen4.csv	Microsoft Offi...	
event_count_time_wise_gen5.csv	Microsoft Offi...	
event_count_time_wise_gen6.csv	Microsoft Offi...	
event_count_time_wise_gen7.csv	Microsoft Offi...	
README.txt	Text Document	

Fig. 3-3 ***Screen shot showing the downloaded CSV files.***

In an 'event_count_generation_wise.csv' file the count of all 9 types of events at each generation is listed. Generations were listed in the first column while different events were listed in the consecutive columns. Thus, each number represented a particular event at a designated generation. The second file 'event_count_time_wise.csv' accounted for all events counted against time (min). In this file, the first column represents time (mins) importantly only those time points were listed when an event occurs. The remaining CSV files were all generation based, therefore, for each generation the time for each event was listed. For example, in the file 'event_count_time_wise_gen0.csv' the event distribution for generation 0 cells i.e. progenitor cells were listed against time and again only those times where a particular event has occurred. Data from these file packages were subjected to different event-based analysis to understand different aspects of cellular dynamics and the dose dependent effect of the anti-cancer agent topotecan (TPT). So far two simple aims have been achieved. (i) the database construction and subsequent data filtering demonstrated that data sieving serves an important role in selecting lineages with specific descriptors; (ii) the lineage filtering plays an important role in shaping the data mining questions and the hypothesis posed in examining the PD responses to therapeutics agents.

3.3.2 Understanding extent of TPT perturbation on population growth in the temporal domain

Time-to-event given by cumulative event curves provided kinetic fingerprints of cellular phenotypic behaviour in a tumour cell population (Feeney et al. 2003). These time-to-event curves could be considered as a collection of individual cell responses and therefore reflect the population as a whole. The shape of the event curves encompasses cell cycle delay (change in slope), cell cycle arrest (gaps in delivery profiles) and possibly the induction of cell death.

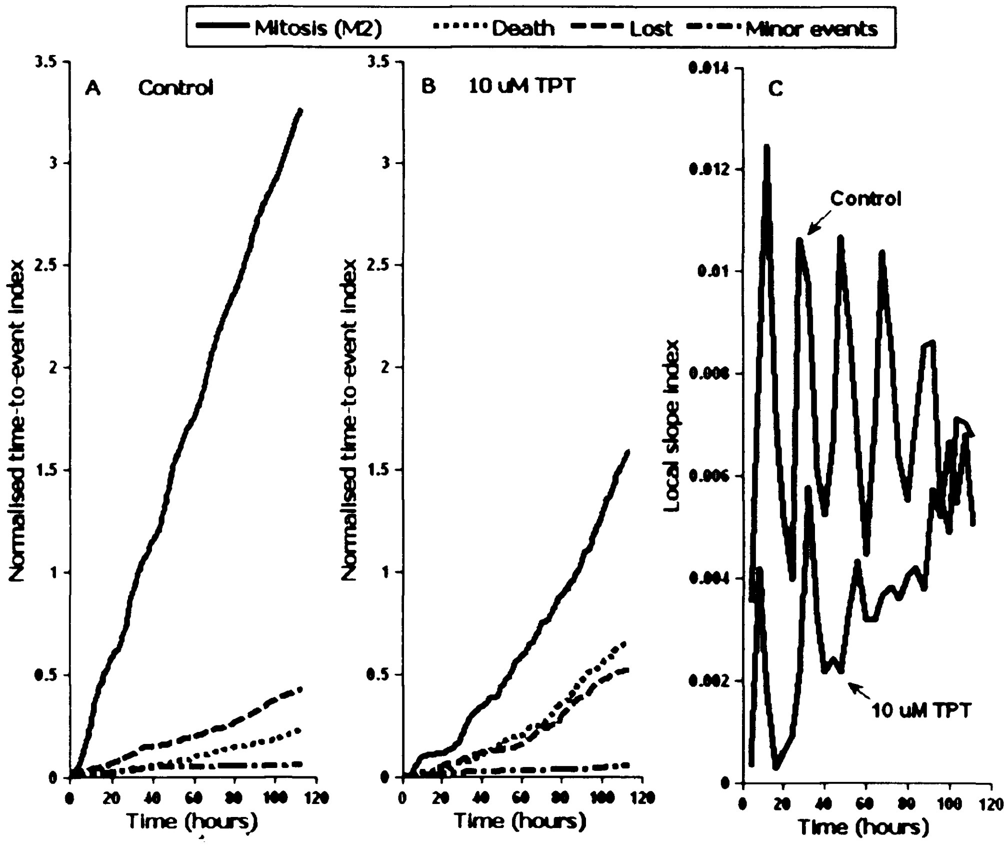


Fig. 3-4 ***Cumulative event curves for U-2 OS cells derived from ProgeniDB.*** *The database was queried to obtain event counts from experimental drug screens of U-2 OS cells in untreated conditions (A) and after a 1hour bolus exposure to 10μM TPT (B) treatment conditions. The tagged events in this query included mitosis (solid), cell death (dotted), lost (dashed) and all other active events (dash-dot; such as polyploidy, refused, mitosis 3 and mitosis 4). The event curves were subjected to a continuous local normalization filter ensuring an adaptation to the concomitant increase in cell number. Local slope changes (C) (4 hour time bin) plotted over time to demonstrate elements of population dynamics, synchronization, and inter-mitotic perturbations. The local slope was calculated from the mitotic events, depicting oscillatory behaviour during normal growth and the consequences of drug perturbation. Continuous normalization filter: The experimental acquisition time interval (tv) was 15 minutes (tv = 15) and therefore the value of N was updated every 15 minutes using the formula described in Eq. 2. The number of living cells, N, at the start of the experiment (t = 0) equals the number of progenitor cells present under the specified condition (in this example, the control condition has 156 progenitor cells and the 10 μM TPT condition has 201 progenitor cells). From this time point, as the population starts to proliferate, N is recalculated as follows:*

$$N_t = N_{t-tv} + \sum Mitosis_{t} - \sum(Death + Lost)_t \qquad \textit{Eq. 2}$$

The value of N at time t was used to normalize each event curve, e.g. for mitosis

$$M2_t / N_t \qquad \textit{Eq. 3}$$

Local slope calculation: To determine local changes in the rate of event delivery there was a requirement to calculate the local slope, this calculation was binned over 4 hour time windows providing sufficient sub cell cycle resolution.

$$Slope = \frac{\Delta y}{\Delta t} \qquad \textit{Eq. 4}$$

Where Δy = value of y_t - value of y_{t-4} and Δt = 4 hours.

ProgeniDB was queried with relevant filters to select a constrained set of lineages, representing a specific drug treatment regimen. In each instance within the downloaded result files, the 'event_count_time.csv' file was used to calculate the cumulated events (mitosis, death, lost, and all other active events), a normalization step was added to provide a continuous adjustment for the number of cells present in the population. Previously a normalization filter was implemented with relation to the original number of progenitor cells (Marquez et al. 2003), this approach is effective for short duration screens (12-36 hours). However, for long term screening (> 48 h) a continuous normalization approach was implemented where the cumulative events at each time point were adjusted for concomitant increase in cell number and hence event potential (total number of cells) at a given time point. The normalized event curves in the control conditions (figure 3-4 *A*) showed a linear ramping of events, with mitosis being the predominant event. The event curves showed a dramatic perturbation effect on the mitotic event curve as a result of a 1 h bolus treatment with 10 μM TPT (figure 3-4 *B*); a triphasic response, with early events arriving, followed by a plateau phase between 8-20 h and then a substantial recovery of mitotic events to approximately half the untreated conditions. This profile represents a complex interplay of pharmacodynamic effects. In this heterogeneous cell population, at any particular time, cells are positioned asynchronously in their cell cycle which is reflected in their event delivery for that time window; (i) the plateau phase represents cells actively replicating DNA during drug treatment, they are unable to contribute to the mitotic curve at the appropriate time; (ii) a drug resistant fraction becomes apparent originating at (>24 hours); (iii) an enhanced rate of cell death occurs (>40 hours), displaying an exponential profile. Figure 3-4 *C* showed a change of the calculated local slopes derived from the mitotic curves in both treatment regimen as the population grows. To reveal the cell cycle driven dynamics within the heterogeneous population, the mitotic event curves were segmented into 4 h time windows and the slope for each binned segment was calculated. In control conditions this yielded an oscillatory pattern, the period of the each oscillation represented the wave of cells delivering to mitosis for each generation in this case five completed cycles; and the sequential decay of the response reflected that population growth tends towards asynchrony. In the treated conditions (10 μM TPT), local perturbations became apparent using this readout approach. It demonstrated the gap or hole in the first cycle (at > 8 h) and a slow ramping up of the rate of mitotic delivery to approximately half that of the control counterpart (from >24 h to end). Interestingly this fraction which was termed as the drug resistant fraction recovered non-synchronously (no obvious oscillatory component), demonstrating a severely perturbed tumour population. The conversion of a linear temporal response to an oscillatory or frequency response by translating via local temporal texture analysis provides a route for mathematical modelling; and the incorporation of wavelet analysis to enable pharmacodynamic fingerprinting in the context of drug profiling.

3.3.3 Understanding the consequences of TPT action on sequential generations of population growth

ProgeniDB provided a unique opportunity to extract event-based analyses from the perspective of a lineage structure i.e. based on generation as opposed to time per se. Data analysis and visualization output using the downloaded 'event_count_generation.csv' file in the same drug treatment conditions as above showed the effect of the drug from a generation perspective.

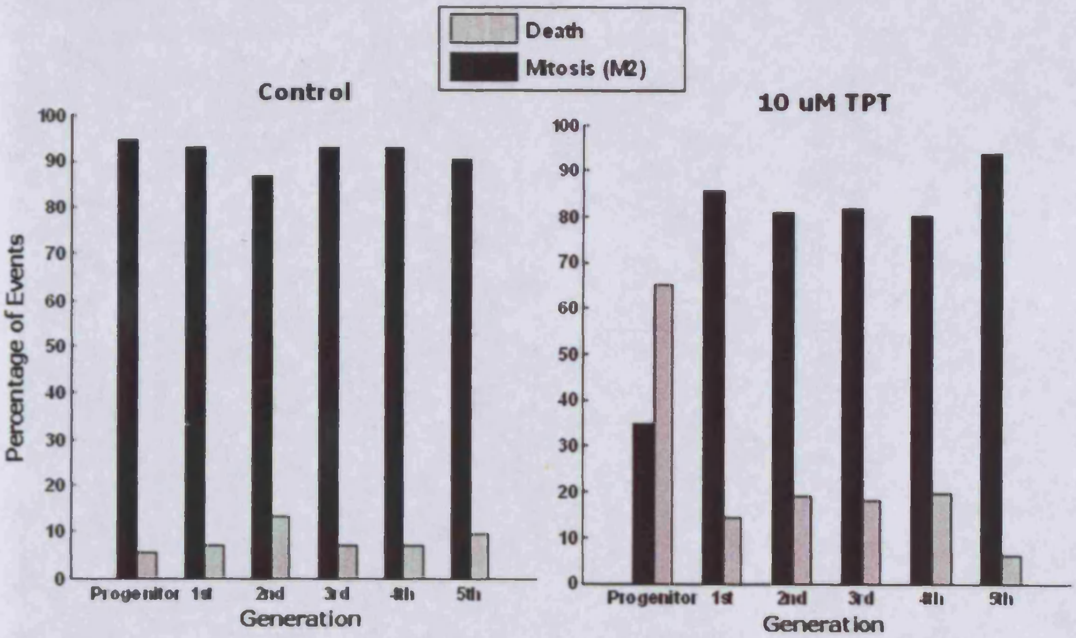


Fig. 3-5 ***Generation derived event analysis.*** *Simple histogram of percentages of mitotic and cell death events calculated for each generation in control and in treated condition.*

The percentage distribution of mitosis and cell death was determined on a generation-basis. In control conditions the level of mitosis was on an average 10-fold greater than the level of cell death over all generations. After treatment, the level of cell death was 2-fold greater than the level of mitosis in the progenitor generation. This effect reverts to predominantly mitosis with an average 4-fold ratio over cell death for successive generations. At a simple level, this visualization provides a means for determining the generation specific cell responses to drugs specifically the behaviour of the resistant population.

3.3.4 Simple data visualization to compare signatures of drug action

To compare the pharmacodynamic response of a treated population with a control population on a generation basis and on a time scale basis, the data from figure 3-4 and 3-5 was reconfigured to provide the following figure.

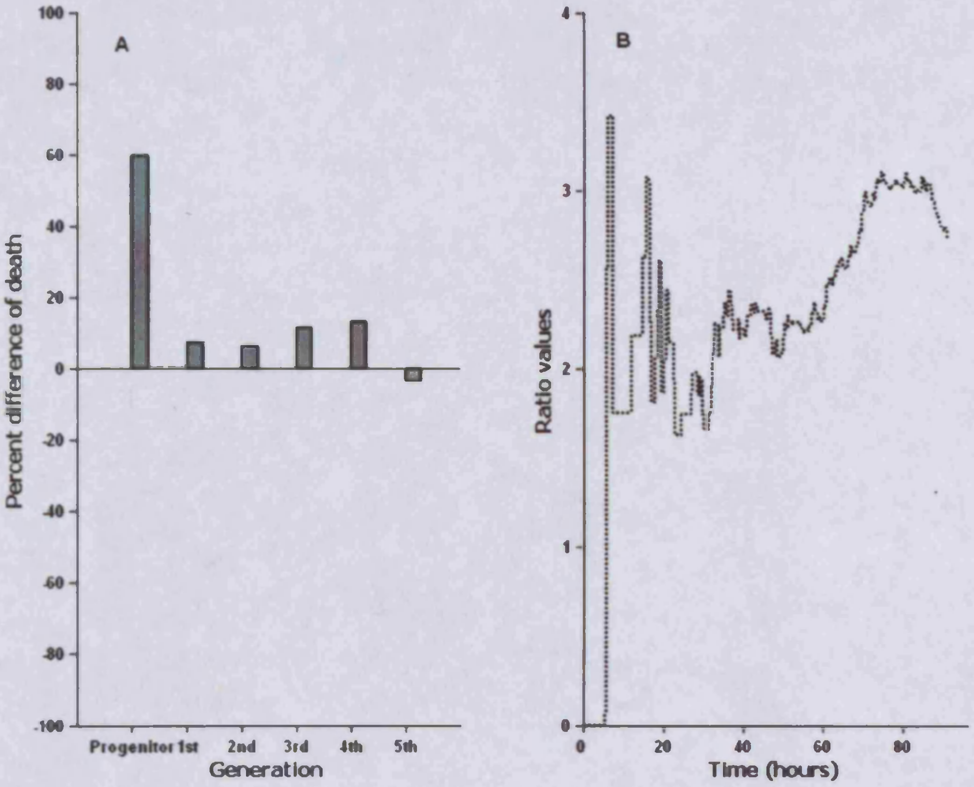


Fig. 3-6 Comparing pharmacodynamic (PD) responses in context of cell death to drug perturbations. (A) Histogram depicting percentage difference of death in different generations. In each generation, percent difference for cell death was calculated by subtracting the control values from the drug treatment (10μM) values of (figure 3-5) (i.e., $\%D_{treated}$ - $\%D_{control}$). (B) Using data from figure 3-4 A and B, ratio of $D_{treated}$ over $D_{control}$ at each time point was plotted against time.

Histogram plot (figure 3-6 A) showing the difference for cell death in treated conditions (10 μM TPT) compared to control conditions was derived from data shown previously. The plot showed for TPT treated condition, that the percentage of death was always higher (i.e. positive) up to and including the forth generation. This revealed that while the response reduced over the consecutive generations the addition of TPT enhanced cell death throughout the integrated population lineage. Figure 3-6 B showed a continuous ratio plot for cell death over time, taken from figure 3-4. Cell death is considered as the critical events in this particular analysis. A ratio value of 1 indicated that the time-to-event index was equivalent in both conditions, while above 1 indicated a higher rate in drug-treated conditions and vice versa. The ratio for cell death stabilized after 20 h and slowly ramped up to a plateau mean value of 3.0 (SD ±0.73), therefore cell death originates from the progenitor cells, and is detected during the latter stages of the experimental period.

3.3.5 Querying the acute effects of Topotecan on the progenitor population to decipher the cell cycle origin of drug resistance

In order to investigate the dose-dependent acute effect on progenitor cells only, ProgeniDB was re-queried for control, 1 μM TPT and 10 μM TPT conditions and, in each instance, data downloaded from the 'event_count_time_gen0.csv' was analyzed. These files contain the time-to-event counts for all nine tagged events for progenitor cells only.

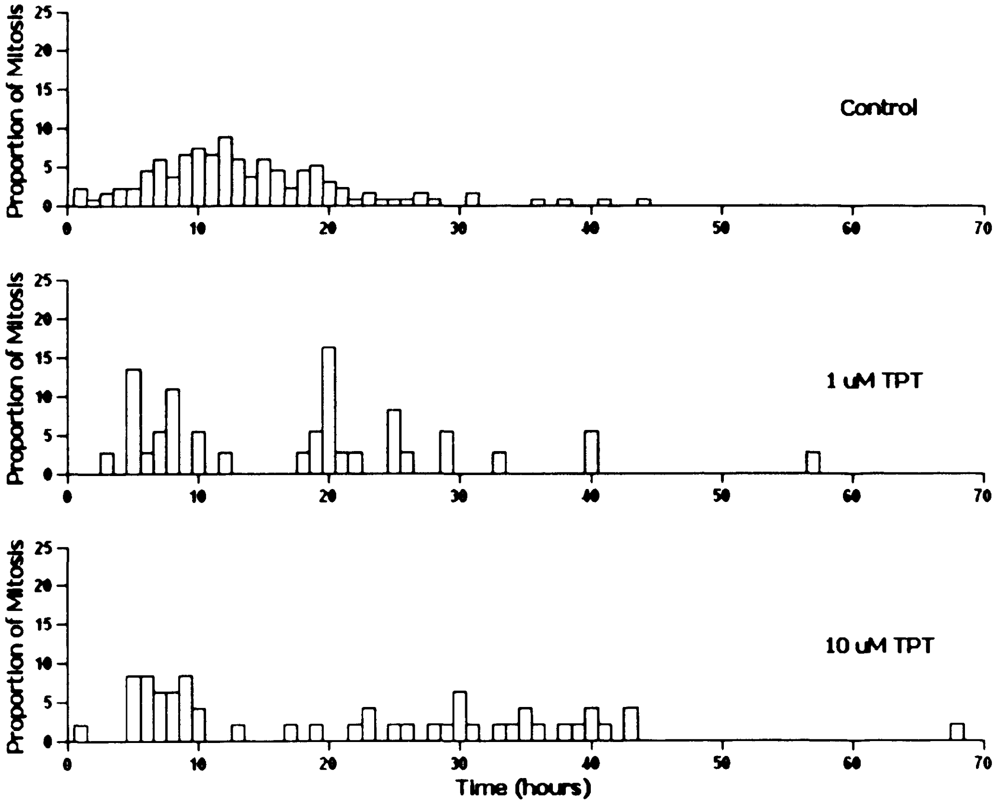


Fig. 3-7 Distribution of mitotic proportion (in one hour time bins) for progenitor cells at three TPT doses (Control, 1 μM and 10 μM). For each condition, the proportion of mitosis at every hour was calculated using the following formula:

$$P_h = \frac{M2_h}{\sum M2_{0\ldots70}} \times 100 \qquad \textit{Eq. 5}$$

Where M2 is the number of mitosis during that time bin hour h and P is the proportion.

The plots showed, that in control conditions within 24 h, 92% of the total progenitor cells have undergone mitosis, while for 1 μM TPT treated and 10 μM TPT treated conditions the value was 73% and 56% respectively, thus the spread or smearing of the distribution indicated the dose-dependent global cell cycle delay effects. Interestingly the shape of the distribution revealed underlying cellular dynamics. The mitotic event nadir seen previously (figure 3-7 *B*) between 8 to 18 h also occurred at 1μM TPT (with both doses operating at 5% of the total progenitor cells delivering during this time zone). However, at 1 μM TPT either side of the 'hole' (the duration of the nadir) a mini-mitotic surge occurred while for 10 μM TPT the response was flat and of a very long duration. To interrogate the impact of the drug on cells delivering to mitosis within specific time windows the histogram plot was partitioned into three zones. In essence these corresponded to cells delivering to mitosis from different origins in the cell cycle. In other words cells delivering to mitosis within the first 0 - 10 h zone could be considered to be in G2 during the drug treatment, S-phase if time to mitosis occurred between 11-18 h, and G1 if delivery to mitosis occurred during the final temporal zone (Errington et al. 2006; Feeney et al. 2003). The dose-dependent breakdown into these three time zones were shown in the following table.

Table 3-1 Percentages of mitosis of progenitor cells in three different time sections.

	Control	1 μM TPT	10 μM TPT
0-10 Hours	**37 %**	**41 %**	**44 %**
11-18 Hours	**42 %**	**5 %**	**6 %**
19-70 Hours	**21 %**	**54 %**	**52 %**

The results were consistent with the previous findings that TPT is an S-phase specific drug and that cells in G1 and G2 represent the potential source of a drug resistant fraction (Feeney et al. 2003; Pommier 2006). Late cell cycle (G2) effects could be dissected in detail from the event analysis data and previous studies have identified the G2 originating fraction within the original population, as cells that deliver to mitosis before the S-phase marked cells (Marquez et al. 2004; Smith et al. 2007a). The first cohort of cells (0-10 h) to deliver to mitosis was considered to be in G2 during the drug treatment. ProgeniDB has provided the opportunity to filter and select a sub-set of cell lineages based on progenitor cell time-to-mitosis. A data query was performed to only address these lineages. For the 10 μM TPT condition, this represented 21 lineages or 10% of the possible lineages (figure 3-8 A)

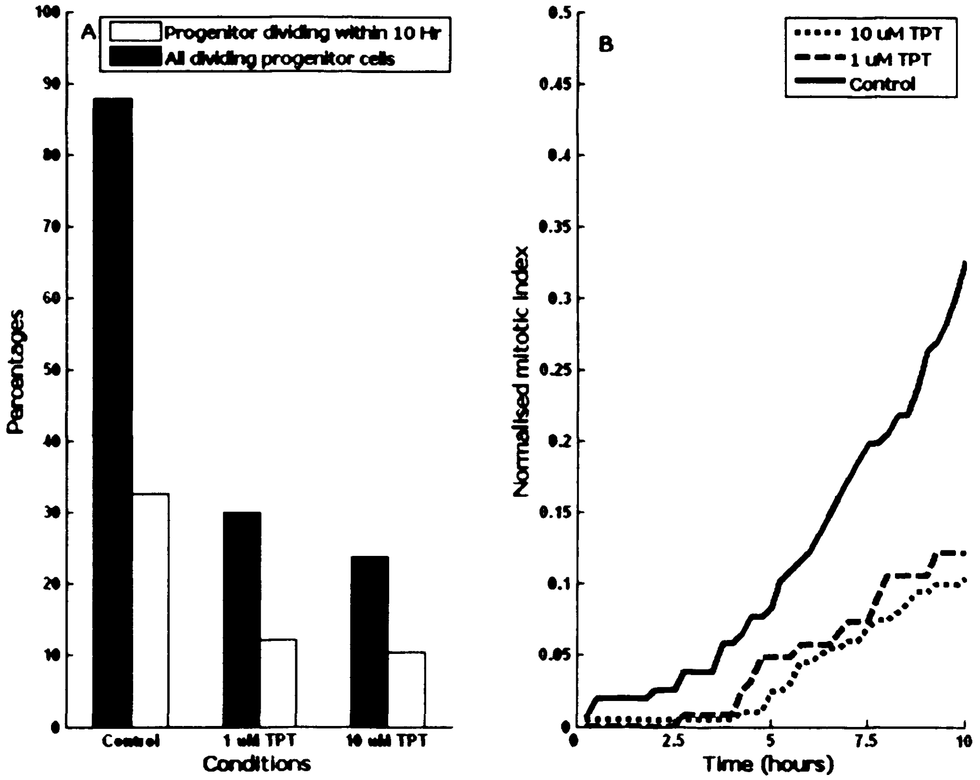


Fig. 3-8 (A) In three different experimental conditions, the percentage of progenitor cells that deliver to mitosis and a subpopulation of progenitor cells that deliver to mitosis within 10 hours of start of the experiment. (B) Normalized rate of mitosis of progenitor cells divided within 10 hours of start of the experiment.

The mitotic event curve for each corresponding G2 fraction of progenitor cells was normalized with respect to the total number of progenitor cells included in the assay for that condition. The addition of the drug abrogated the delivery of this fraction of cells to a similar extent at both drug doses. Therefore it can be concluded that the late G2-checkpoint is induced by TPT.

3.4 Concluding remarks

The ProgeniDB database introduced a new approach to access information on dynamic cell behaviour. Fundamental to the ProgeniDB concept is that the encoding process encapsulates critical features of cell-cell heterogeneity and time-dependent events. The multi-level descriptors and parameters attributed to each node within the resultant cell lineage maps provided a unique framework for applying bioinformatics-like query algorithms such as those used for genomic databases. The lineage map importantly provides generation and cell functional layer upon which other information can be linked, such as proteomic and genomic expression data.

It is important to note that all these results shown in previous sections utilizing the CSV files downloaded by querying ProgeniDB was also produced by querying the LDB. Since both database used same data but in different format the results were identical (data not shown). Indeed this process validates the design and implementation of ProgeniDB, however the mining and analysis on LDB was not performed with the intention to validate ProgeniDB rather these and other data mining (vertical and horizontal) and associated analysis were performed on LDB only because of the simplicity of tab delimited text file. Moreover the nature of this research was hypothesis-driven data mining that demands a 'trail and error' based approach that again supports text file format contrary to MySQL format. In terms of work flow, once confidence was developed on a particular type of mining and analysis (e.g. event distribution) process, the associated mining and analytical algorithms were then implemented in MySQL format as in the case of ProgeniDB.

The future improvement of ProgeniDB depends on the mining and analytical complexity that can be achieved on LDB data. The next chapter illustrates a wide range of analyses that were performed on the LDB data to illustrate the level of complexity and understanding these encoded data can provide. These mining and analytical features will be incorporated to ProgeniDB in due course but present endeavour can be deemed as a proof of concept that demonstrates the operational reality as well as value of establishing such prototype database. ProgeniDB will not be a standalone database rather will be part of an unique e-science (Fox et al. 2003; Hey and Trefethen 2003) endeavour – CyMART, which will be discussed in detail in the concluding chapter. ProgeniDB and other future databases that stores cellular data acquired through different microscopic technology (e.g. for fluorescence microscopy, FluorDB will be developed) will be incorporated in CyMART. These suite of databases will be the access point to retrieve cellular data pertinent to different research areas like cancer, wound healing etc, the viewpoint of such databases will be like that of NCBI or EBI, a web portal with arrays of databases like Entrez Gene (see http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene) with mining facilities, analytical and visualization tools like BLAST (Altschul et al. 1990), Clustal W (Pearson 1990; Pearson and Lipman 1988). Construction and update of ProgeniDB and other cellular databases will also be achieved in a collaborative manner like NCBI, the process of encoded data contribution was outlined thematically earlier in this chapter.

Chapter 4: Lineage maps: revealing interactions and relationships

The concept behind this investigation is that the progeny tree generated from progenitor cells provides the structure of the lineage map where access to data depict nodes of cellular behaviour that can be mined and analyzed from different perspectives. Cellular event analysis and distributions have remained the predominant form of analysis in cell based assays, where the basis was event counting (both manually and automatically). Previous chapters have illustrated access to such event information acquired from ProgeniDB (Khan et al. 2007). The information has been interrogated from different perspectives, in relation to time and generation which was achieved by segmenting lineages vertically (discussed earlier in chapter 2). However, these analyses did not exploit the core strength of the lineage data format, which is the inter-nodal relationship, where these aforementioned parameters like event type, IMT etc. could be analyzed within a pedigree structure. Lineage maps represent all events as nodes which are inter-connected by branches, if both nodes of a branch represent mitosis, branch length represents the inter mitotic time (IMT), in other words the complete cell cycle time. IMT would be considered the most significant parameter in these type of analysis and the variability of IMT, monitored from generation to generation (pedigree structure) after an insult or perturbation is a good indicator of the time-dependent action of a drug. Specifically, the pattern of the delay could provide a dose and drug class-dependent signature that can be mined from the lineage database (LDB). Moreover as these tracks represent direct lines of descent, behaviour of later progeny can be attributed to previous generations particularly the progenitor cell (the first cell or the root of the lineage). Relating later progeny behaviour with progenitor cell attributes, provides knowledge concerning the origin of drug resistance and facilitates the need to develop predictive algorithms and models of pharmacodynamic (PD) responses.

It is important to stress that the values of individual IMT or other inter-nodal data were not encoded within the text files and therefore were not stored in the LDB (folder structured database where encoded text files were archived initially) rather they were extracted from the lineage data 'on-the-fly' basis, during the data mining process by calculating the time difference between two (branch length) successive mitosis. From the LDB, through exploiting this inter-nodal relationship not only IMT but other parameters can be retrieved 'on-the-fly': Inter mitotic displacement (IMD) as described in chapter 2 is one such example. Another such parameter is intrA-Mitotic Time (AMT) which can be measured as the time difference between start and end of each mitosis. Since ProgeniTRAK encodes the start and end point of each event including mitosis, so this parameter can be mined from the encoded lineages which in biological terms represents the end of G2 phase and commitment to mitosis. AMT becomes an important parameter when aspects of the spindle-assembly checkpoint are being considered.

The overall objective of this chapter was to illustrate different data mining and analysis strategies using LDB through exploiting the lineage format. Mining and analysis were executed in two successive stages; the first stage included distribution analysis on parameters that can only can be measured 'on-the-fly'. The second stage included horizontal segmentation, i.e. relationship based analysis to explore drug induced pedigree behaviour, origin of drug resistance and origin of population heterogeneity.

4.1 AMT – representing M phase variability

Many cellular events, particularly mitosis have well defined start and end points, and this coincides with distinct morphological cellular changes; within this bioinformatics environment this time difference was designated AMT which was measured from the LDB 'on-the-fly' basis. The time window or phase (representing the M phase) of the cell cycle is important as various drugs interfere with this process to cause mitotic stalling and arrest, for example the Paclitaxels a mitotic inhibitor used in cancer chemotherapy Taxol® (Kraut et al. 1996). Taxol® stabilizes microtubules and the consequences are the capture and delay cells in the mitotic state as they stall at the spindle-assembly checkpoint (Abal et al. 2003; Allman et al. 2003). The extent of this delay is an indication of the PD response and as ProgeniTRAK encodes the start and end point of each event including mitosis, this aspect of a PD response can be measured and presented in a comparative manner within a drug screen.

The impact of Taxol® and TPT (two well studied anti-cancer agents) on AMT was evaluated, any perturbation of this phase of cell cycle was usually observed as an elongation of AMT value and a reflection of the spindle-assembly checkpoint in action.

Using timelapse imaging, it has been previously revealed that the dynamics of taxol-related agents in altering mitosis are complex (Rosa et al. 2006). Mitosis can be prolonged but cells are still capable of dividing to produce either two or three cells (tripolar mitosis), thus explaining a sub G1 peak as aneuploidy rather than apoptosis. Furthermore, some cells are able to fuse back and then progress to mitosis, frequently producing three cells again before becoming arrested in the next cell-cycle interphase (Demidenko et al. 2008). The effect of drug on AMT was measured for (U-2 OS) progenitor cells that were exposed to 1 μM TPT, 10 μM TPT and 5 nM Taxol® as well as in control condition. The aim was to seek long term stalling of mitosis, hence AMT values were binned according to their nearest hour and for each hour the percentage or proportion was calculated using the following formula:

$$P_t = \frac{N_t}{\sum N} \times 100 \qquad \text{Eq. 6}$$

Where P is the proportion at time t^{th} hour and N is the number of mitotic events where mitotic events include mitosis 2 (M2), mitosis 3 (M3), mitosis 4 (M4), polyploidy (P) and re-fused (R) events.

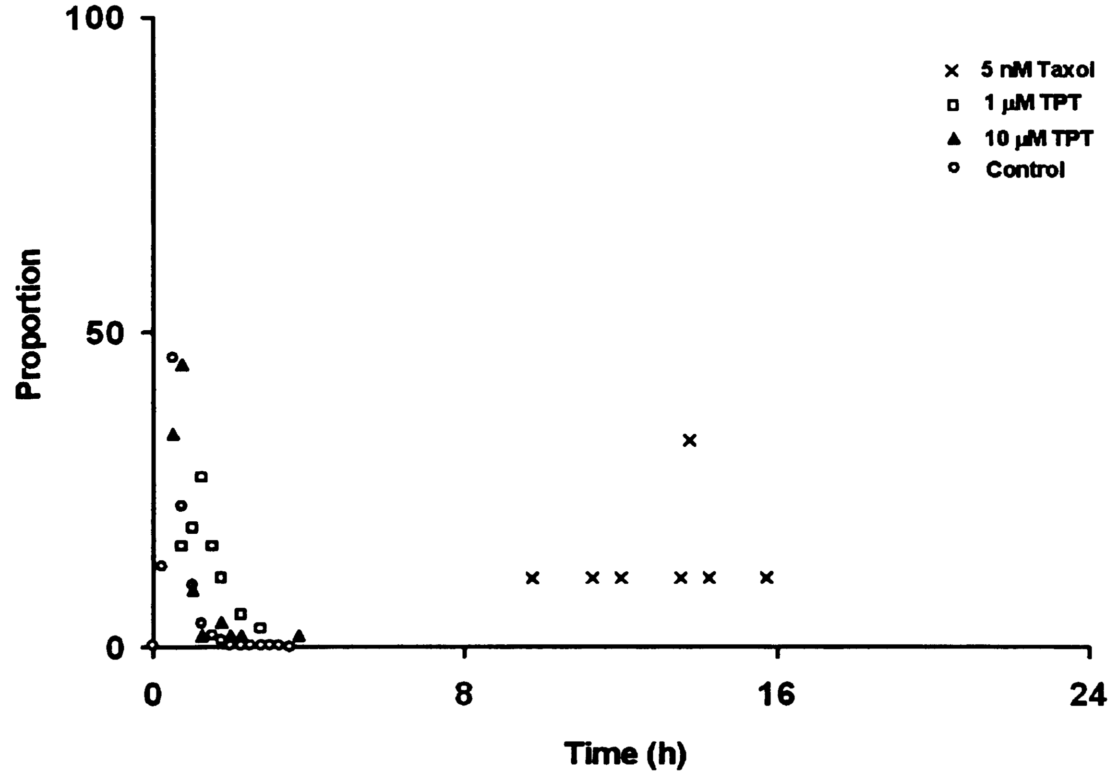


Fig. 4-1 Effect of Taxol® and topotecan (TPT) on the AMT of progenitor cells.

The distribution of AMT in U-2 OS progenitor cells showed a marked elongation with 5 nM Taxol ®, while with TPT no visible effect was observed. In addition to an elongation of AMT it was also observed that all mitotic events resulted in a polyploidy or re-fused event outcome (data not shown), this indicated a failed segregation of chromosomes or a furrow-regression event respectively. This multi-dimensional analysis (elongation coupled with event type) demonstrated an exemplar PD response in a comparative manner that is necessary for understanding the biological action of the drug. More over the PD behaviour from a well known drug i.e. Taxol® can be used as a signature to compare the PD response of new compounds, which in turn would facilitate the process of drug discovery and development. The encoded lineage approach enabled different timelapse experiments to be combined and mined based on the selection criteria (filter) of a single common descriptor, for instance in this example, based on the common cell line (U-2 OS) lineages from three experiments were included for analysis. The first two experiments were identical (with TPT) but the later one (with Taxol ®) was different with respect to experiment duration, sampling interval etc. It is worth noting here that the sampling interval applied during the timelapse acquisition dictates the sensitivity of the AMT assay, current sampling of data in the LDB consists of 15-20 minute intervals while the mitotic event for U-2 OS cells had an AMT duration of ~ 1.5 hours, indicating subtle perturbations would remain undetected. Comparing cellular behaviour in such an incorporated manner would facilitate collaborative research, as independent experiments (performed in different labs) derived lineages can be analysed and interpreted by single analysis.

4.2 Interlinking of cell cycle duration with cellular displacement

Both IMT and IMD are important parameters as they represent the duration of the cell cycle in context with displacement and their variation implies modulation of the cell cycle related events. Assays with TPT on U-2 OS cells (used to encoded lineages) showed that cultures reached confluency, after 100 hours particularly in control conditions (see Appendix VI for image sequence) – affecting cell movement. The question was whether such confluency has an impact on IMT, as well as the inherent mobility dynamics of cells. To determine the co-effect of confluency, the mean IMT and IMD in successive generations was measured and presented in the following plot.

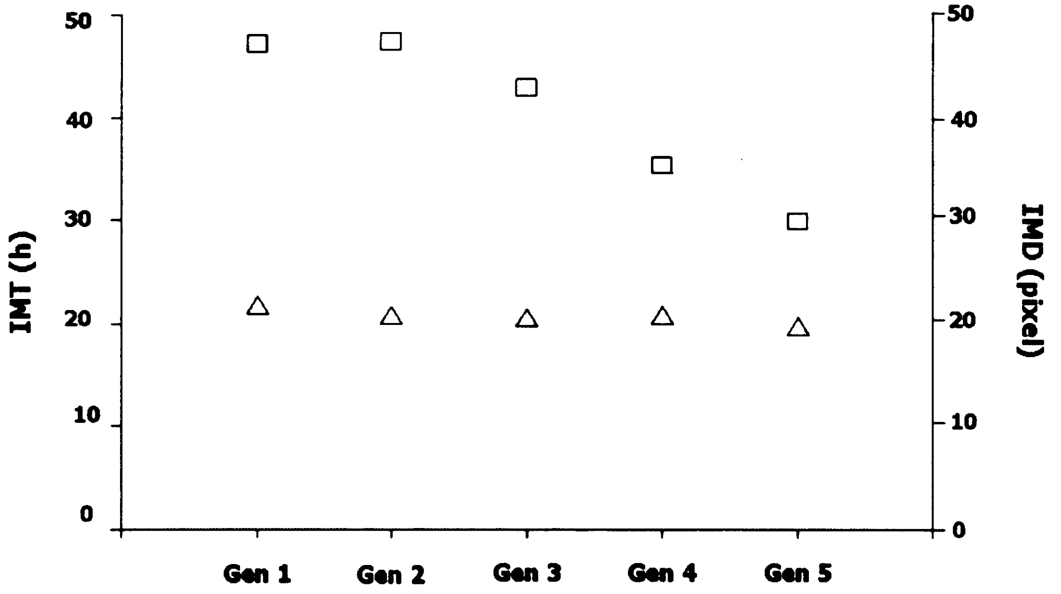


Fig. 4-2 ***Stability of the IMT and IMD over five generations.*** *The mean IMT (triangle) and IMD (square) values for five generations were plotted in a two axis graph, it was revealed that the IMD decreased 30.54% over a five generation assay period (~ 112 h), however the IMT remained stable during the whole assay period.*

From figure 4-2 it is evident that due to crowding the free movement (actual distance moved between two successive mitosis) decreased in generations 4 and 5 but the average IMT remained stable 20.5 h over the course of the entire experiment, indicating that crowding had not imparted upon cell cycle progression. Such an analysis could be an important indicator fo the evaluation of contact inhibition. Further, the analysis validates the assay design providing a standard value for IMT which remains unaffected by the cell crowding. Moreover from a bioinformatics point of view, this result exemplifies the implication of the data format by illustrating the relationship of IMT and IMD with experimental descriptors e.g. dose and generation.

The IMD is a good indicator for measuring the overall rate and directionality of cell movement over time; parameters of great importance for wound healing where a balance of proliferation and motility is required for wound closure. Feedback from biologists confirmed the importance and requirement of the IMD parameter, therefore a visualization GUI was developed where IMDs for a selected lineage can be viewed and quantified. The workflow was designed as follows: the user selected a particular field of view through a sequential selection process resembling lineage encoding process - experiment selection followed by well selection which led to a field of view (FV) selection. Within the selected FV visualisation of the progenitor cells (green dots in figure 4-3) were displayed. Each progenitor cell represents a single lineage, therefore when user selected a particular progenitor cell or green dot, all IMDs attributed to each node of the lineage were drawn as lines centred for the progenitor cell coordinate origin (starburst effect). Additionally, the lines were colour coded according to the generation of IMD they represent.

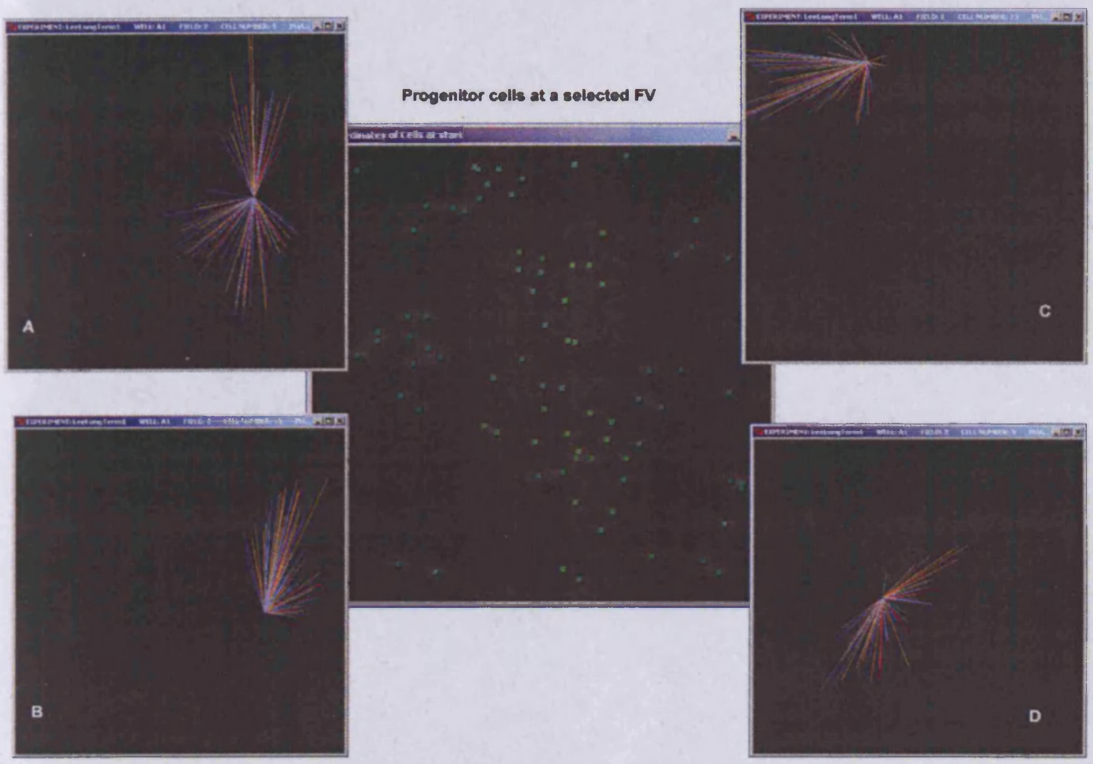


Fig. 4-3 ***IMDs in different generations.*** *Four progenitor cells were selected from the middle screen shot where each green dot represents a progenitor cell or root of the lineage. The progeny evolved from each progenitor cells (A-D) has associated IMD cluster which were depicted as lines centring to the origin (x,y coordinates of the progenitor cell).*

Since these type of analyses provided both quantitative measurement and a visual representation, they naturally established their importance in simple and well established assays where colony formation pattern (size and number and potential sectoring) were parameterized. A standard method for assaying a population of tumour cells for drug sensitivity involved plating an equal number of cells with and without drug and then comparing the two groups on the basis of the number of colonies grown to 50 cells or more. Routinely this required a typical plating of 20 cells in a 60-mm tissue culture dish, incubation for 4-14 days, followed by a staining procedure for colony detection, counting and size determination (Mather and Roberts 1998). There were semi-automated platforms for extracting counts and analysis including mean colony diameter, area, density and distance to nearest neighbour, as well as colony size distribution (e.g. ColCount™, from Oxford Optronix). However the limitation of this classic approach was that cells have to be sparsely plated to identify the physical colony boundary, and a minimal seeding density is not appropriate for many cell types and normal plating conditions maximizes the growth of progenitors of interest (and their clonal progeny). Furthermore it may be informative to take into account clonal heterogeneity of growth rate (Kuczek and Axelrod 1987). The single-cell lineage-derived colony parameters described in this chapter form the basis for a multi-dimensional colony assay, incorporating the limitations of the classical colony assay approach and can be implemented using linked IMT-IMD parameters to embed growth rates.

It is important to note that cell directionality here represents the overall movement between lineage nodes (as in ProgeniTRAK data were encoded only when an event occurred). Even with this restriction, IMD is an important parameter as it relates cell movement and with cellular proliferation and other experimental descriptors - generation and time, this matrix of information provides the global analysis required for wound healing and metastasis (detail discussion in chapter 7).

4.3 Exploring cell cycle duration patterns: measuring the dissemination of perturbation effects through successive generations

The primary focus of the current chapter in the Intra Mitotic Time (IMT) (technically the inter-nodal distance) and the specific hypothesis was that this bioinformatics-derived parameter enabled the biologist to dissect out the mechanisms that underpin the dynamics of progenitor cell lineages. Furthermore this parameter represented the global analysis of the cell cycle engine. The first premise was that interrogating the IMT parameter using a multi-dimensional approach was now possible because of the lineage format. Single cell lineage tracks (i.e. direct lines of descent) provide temporal tracks for cell division and cell death timing, forming a bifurcation map with nodes, branches and cul-de-sacs. The duration of the IMT, monitored from generation-to-generation, after an insult, is a good indicator of the time dependent perturbation of cell cycle traverse in response to cell cycle delay-inducing drugs. The pattern of this delay from generation-to-generation could provide a dose and drug class-dependent signature. Lineage data encoded through ProgeniTRAK provides the unique opportunity to explore the existence of such signatures, which from a cancer research perspective has two clear benefits – first the pattern of PD response attributed to resistance sub-populations can be identified and second from the roots of these resistance tracks, the origin of resistance can be identified. Even though in the previous analysis it was shown that cell cycle positioning of the progenitor cells was an indicator for survival at high doses (1 µM and 10 µM TPT), it was shown that progenitor cells either positioned in G2 or G1 had higher survival percentages than cells positioned in S-phase (see figure 3-7), but from that analysis it could not be revealed the fate or behaviour of the successive progeny of this apparent resistant progenitor cell sub-population. In order to achieve this objective a reference or standard IMT from the control experimental condition was required against which other outcomes of cell cycle delay or early etc. could be scored. The approach is analogous with PAM or BLOSUM scoring matrices which specify the similarity or the distance of replacing one protein residue/base by another, based on the theory of evolution (Dayhoff 1979; Henikoff and Henikoff 1992). These scoring matrices are used in BLAST to align unknown protein sequence with well annotated protein sequence in the database. The previous analysis with control population showed that the mean IMT remained almost constant over five generations (see figure 4-2), however the true distribution of the IMT has not been comprehensively evaluated through Johnson curve fit (Johnson 1949). *A mini-project was devised to interrogate this aspect further, (see Appendix VII) undertaken as a Masters project by Chris Headley.* The critical outcome from the study showed that the distribution followed a normal distribution and therefore the IMT control reference was designated as 20.5 h ± 5.02 SD.

It was decided that the pattern of IMTs for each track should include successive three IMTs because for treated condition (for this analysis with anti cancer drug TPT) if the progeny of cells continued to deliver even after three generations – at least from the biological perspective these cells were designated as the resistant population. These would comprise the non-S-phase fraction as *a priori* knowledge indicated that S-phase specific progenitor cells did not deliver to mitosis at the two doses selected (1 µM and 10 µM TPT). The mean IMT in control population for U-2 OS cell line was found to be 20.5 ± 5.02 SD. setting this IMT control reference value to 20.5 h or 1230 min a simple scoring schema was formulated.

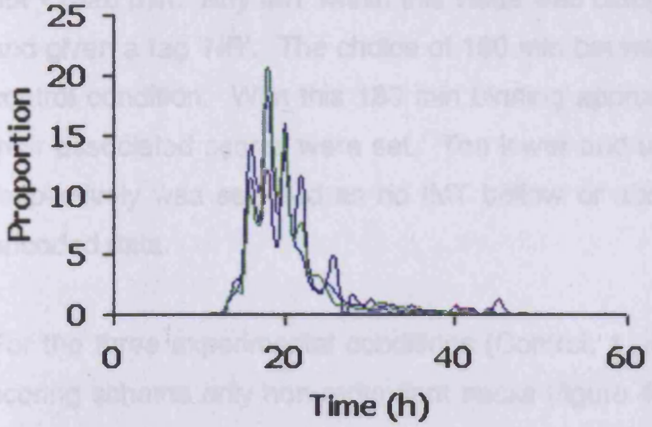


Fig. 4-4 ***IMT distribution of over three generations.*** *The mean IMT from generation 1 to 3 was 21.20 h, 20.01 h and 20.26 h respectively depicted by green, blue and black lines respectively.*

Variation of IMT or cell cycle duration was categorised into 7 distinct bins, where each bin has a time range of 180 min. If the IMT under consideration falls within a certain bin, that IMT was tagged and scored according to the score as outlined in table 4.1. These seven bins were given a score value form -1 to +1 scale with a 0.33 score interval. This score range (-1 to +1) was selected to set the data within a visualisation range compatible with microarray gene expression visualization technique visualisation used to visualise up and down regulation of gene (Kaushal 2004).

Table 4-1 Tagging and scoring schema for IMTs.

IMT duration (in minutes)	**Category (Tag)**	**Score**
<780	Severely Early (SE)	-1
≥780 but <960	Moderately Early (ME)	-0.66
≥960 but <1140	Lightly Early (LE)	-0.33
≥1140 but <1320	Normal (NR)	0
≥1320 but <1500	Lightly Delay (LD)	0.33
≥1500 but <1680	Moderately Delay (MD)	0.66
≥1680	Severely Delay (SD)	1

Each bin comprised of 180 min range and setting 1230 min as the mid point of the 4th bin as well as the ‘Normal’ category bin, the normal IMT range was calculated to be ≥ 1140 but <1320 min. Any IMT within this value was categorized as normal and was scored 0 and given a tag 'NR'. The choice of 180 min bin was derived from the 5.02 SD value of control condition. With this 180 min binning approach other bins were formulated and their associated scores were set. The lower and upper limit of 780 min and 1680 min respectively was selected as no IMT bellow or above this value was found within the encoded data.

For the three experimental conditions (Control, 1 μM and 10 μM TPT) using the same scoring schema only non-redundant tracks (figure 4-5) with at least 3 consecutive IMTs were selected.

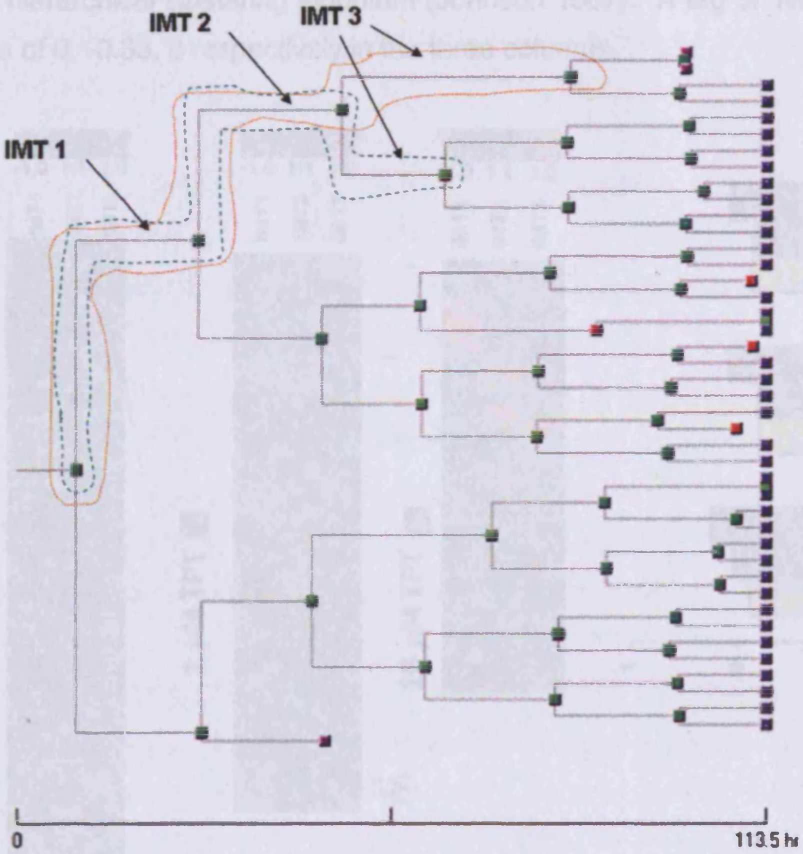


Fig. 4-5 Exemplar lineage where two non-redundant tracks (red and green) up to three consecutive IMTs are highlighted. For these two tracks IMT1 and IMT2 were identical but IMT3 was different.

For each experimental condition, each track was given a searchable nomenclature comprising a 6 letter tag (2 letters for each successive IMT e.g. NRLENR representing a track where IMT1 was 'Normal', IMT2 was 'Lightly Early' and IMT3 was 'Normal') and written to a MS Excel file. For control condition 125 different types of tags were generated, and for 1 μM and 10 μM conditions 52 and 48 types were generated. In order to visualize the tags within a heatmap perspective conventional microarray data analysis software – Genesis ® (Sturn et al. 2002) was used. Even though the tags were ranked according to their prevalence but heatmap view gives a visual representation of the overall behaviour or pattern. In order to achieve this objective, the numeric counterpart (score) for each tag was inputted into Genesis software and was clustered using hierarchical clustering algorithm (Johnson 1967). A tag of 'NRLENR' would have scores of 0, -0.33, 0 respectively in the three columns.

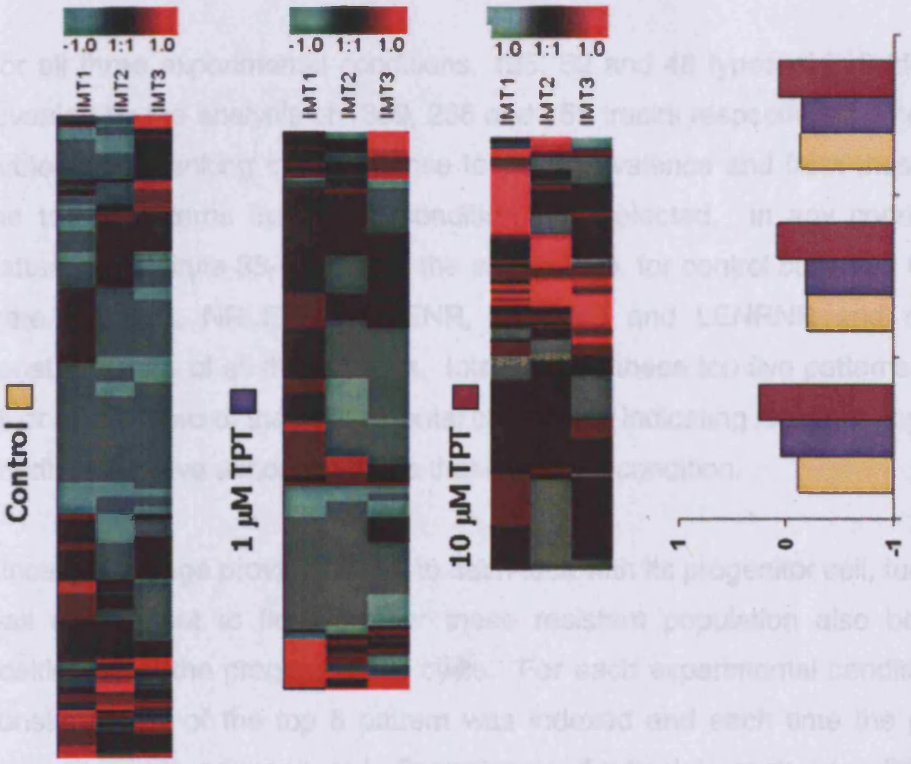


Fig. 4-6 ***Heatmap view of three consecutive IMTs.*** *For each experimental condition, each track of the lineage is represented by a row of the corresponding heatmap and again the three columns in the heatmap represents consecutive IMTs for that track - first column represent IMT1, second IMT2 and third IMT3. If any of the IMT is delayed from the standard IMT, it is coloured as red while early as green. The heatmaps in control, 1 μM and 10 μM TPT represents 1369, 286 and 155 tracks respectively. The average of the score in each column for all three conditions were calculated to indicate the average IMT of the cells at that generation and was presented as bar graph.*

The heatmap (figure 4-6) and the associated clustering were purely for visualization purpose since the underlying algorithm for clustering is specific for microarray analysis which usually has much higher number of genes in the analysis. However from this simple analysis it was apparent that for a high dose (10 μM TPT condition) that the resistant cells showed two distinct delay patterns. The first, where the tracks contained a severe delay specially at IMT1, while in the second pattern cells appear to be completely unaffected by the drug treatment. At the lower dose (1 μM TPT), a small delay occured in the first generation IMT (IMT1), with subsequent recovery. To investigate these findings further, an average of the score for each column was calculated for each condition and was presented in the inset graph. From this simple yet informative analysis a dose dependent behavioural pattern of the resistant population was revealed which also invokes further investigation to ascertain the origin of these resistant sub-fractions.

For all three experimental conditions, 125, 52 and 48 types of individual patterns were revealed by the analysis of 1369, 286 and 155 tracks respectively. These patterns were subjected to ranking in accordance to their prevalence and from these ranked pattern, the top 5 patterns from each condition was selected. In any condition these top 5 patterns constitute 35-40% of all the patterns i.e. for control condition the top 5 patterns were LELELE, NRLELE, NRLENR, LELENR and LENRNR and cumulatively they constitute 34% of all the patterns. Interestingly these top five patterns were common to all or at least two of the experimental conditions, indicating resistant population in treated condition behave almost same as that of control condition.

Since the lineage provided a link to each tack with its progenitor cell, further investigation was carried out to find whether these resistant population also benefited from the positioning of the progenitor cell cycle. For each experimental condition, the track that constitute any of the top 5 pattern was indexed and each time the positioning of the progenitor cell was measured. For example, if a track in control conditions had a pattern LELENR, it was indexed for the analysis and the division time of the progenitor cell of this track was measured. Since a group of tracks belonging to the same lineage, thus will refer to the same progenitor cell; Eventually a distribution of the division time for the progenitor cell was generated that was then presented as a bar graph.

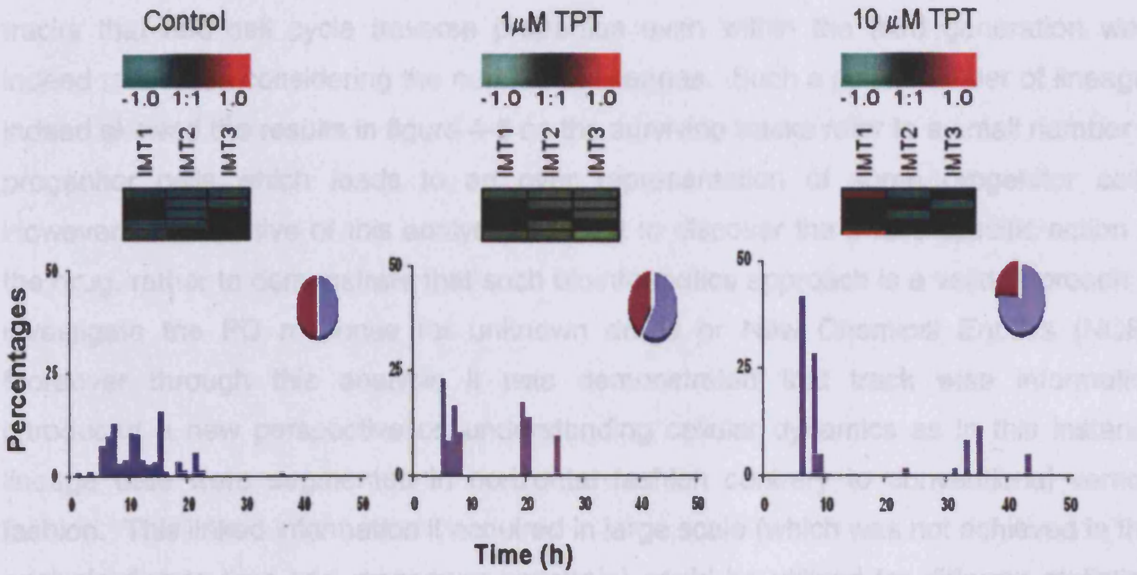


Fig. 4-7 Five most frequently occurring patterns of resistant populations and the distribution of their progenitor origin. Top panel showing the heatmap of the top five patterns in all three conditions while the bottom panel showing the progenitor that are G2 (blue) or non-G2 (brown) origin.

As mitotic time of the progenitor cells retrospectively indicates the cell cycle position at time 0 (i.e. when drug was given) from the bar graph it was evident that for treated condition there were two types of progenitor cells that generated the resistant progeny. The two groups in treated conditions showed a clear gap between 10 to 20 h duration. Having 10 h as the demarcation time point, the progenitor cells could be separated into two groups – progenitor cells that divide within 10 h of start of the experiments were labelled as G2 i.e. they were positioned in G2 of the phase of the cell cycle when the drug was administered and the rest were labelled as non-G2. The chart in the inset shows the percentages of these progenitor cells representing these two groups (figure 4-7). For control both G2 and non-G2 progenitor cells contributed equally to produce the most frequently occurring tracks, however in treated conditions the predominance of G2 progenitor cells were evident with 10 μM being more predominant than 1 μM TPT. This result implies that with higher dose the resistant population predominantly evolved from progenitors that were positioned in G2 of the cell cycle when the drug was administered.

From a statistical view point these analyses were not robust enough to establish any PD response related hypothesis and the prime reason being the small number of tracks available for analysis, specially in treated conditions. At a higher dose the number of tracks that had cell cycle traverse properties even within the third generation were indeed rare when considering the number of lineages. Such a small number of lineages indeed skewed the results in figure 4-6 as the surviving tracks refer to a small number of progenitor cells which leads to an over representation of some progenitor cells. However the objective of this analysis was not to discover the phase specific action of the drug, rather to demonstrate that such bioinformatics approach is a valid approach to investigate the PD response for unknown drugs or New Chemical Entities (NCE). Moreover through this analysis it was demonstrated that track wise information introduced a new perspective on understanding cellular dynamics as in this instance lineage data were segmented in horizontal fashion contrary to conventional vertical fashion. This linked information if acquired in large scale (which was not achieved in this analysis due to time and manpower constrain) could be utilized for different statistical analysis such as cluster analysis and through these metadata level analysis more complex biological hypothesis could be addressed which in turn would augment our understanding of cellular dynamics (temporal domain) and the extracting of *in silico* drug response.

4.4 Cell cycle positioning of progenitor cell and measuring its influence on behaviour of cells on successive generations

Statistical work of this section was carried by Dr. Valentina Moskvina (Biostatistics and Bioinformatics Unit, Cardiff University)

Previous lineage based analysis (section 4.2 and 4.3) has shown that the cell cycle position of the progenitor cell was a determinate for a resistant population to evolve. Not surprising since topotecan forms a ternary complex with topoisomerase-DNA, consequently collisions of DNA replication forks (during replication in S-phase) or progressing RNA polymerase molecules (during transcription in all the cell cycle) cause double strand breaks (Pommier et al. 2004) and evoke a DNA repair cascade. TPT is therefore considered an S-phase specific agent, with differential sensitivity on G1 and G2 cells (Feeney et al. 2003). Furthermore, previous analysis (section 3.3.5) has indicated that from 0-10 hours in a timelapse sequences the cells that deliver to mitosis during this time window could be considered as the G2 progenitor cells, after this time cells are classified as non-G2. Within the non-G2 group it was evident from both (figure 3-7 and figure 4-7) analysis that almost no progenitor cell delivered to mitosis between a ~ 10 – 18 h window at different TPT doses. Again this refers to the well established understanding that S phase specific cells undergo apoptosis or become arrested due to the action of this anti-cancer drug (Feeney et al. 2003; Huang et al. 2003), in addition the late arriving progenitor cells (cell delivering to mitosis after ~ 18 h) could be classified as G1 cells, thus two distinctive sub-populations became evident in the TPT treated conditions which was primarily attributed to the asynchronous but cell cycle dependent delivery to mitosis by the progenitor generation. The impact of such an asynchronous mitotic delivery on successive generations has not been investigated in previous analysis; however, the encoded lineages provided the opportunity to investigate the sub-population behaviour in successive generations and the principal influences. Based on the *a priory* knowledge that TPT is an S-phase (Feeney et al. 2003; Huang et al. 2003) specific drug (particularly at the high doses) and resistant progeny evolve from progenitor cells that were positioned to either G2 or G1 phase of the cell cycle when the drug was administered, an investigative data mining and statistical analysis was undertaken to measure the sub-population behaviour in the successive generations.

Lineages derived from each experimental condition (Control, 1 μM TPT and 10 μM TPT) were selected and for each condition the distribution of mitotic delivery in three successive generations were analyzed. In order to measure the mitotic delivery distribution for a particular generation under a particular experimental condition, all mitotic events at that generation were first accounted and then cells contributing to mitosis were categorized to either G2 or non-G2 origin according to the mitotic delivery time of the progenitor cell of the lineage in which cell under consideration belonged to. Once categorized, the percentages of mitosis were quantified using a 1 hour bin, i.e. percentages were calculated at each hour 0 to 112 h (experimental duration) using the following formula:

$$P_t = \frac{M2_t}{\sum M2_{gn}} \times 100 \qquad \text{Eq. 7}$$

Where, P_t is the percentages of mitosis for a particular category, $M2_t$ is the number of successful mitosis occurred at t^{th} h with cells belonging to that category. $\sum M2_{gn}$ is the summation of all mitotic events (both G2 and non-G2 category) that have occurred in generation (gn) under consideration.

The percentages for three successive generations in all three experimental conditions generated nine graphs depicting the distribution of mitosis (figure 4-8).

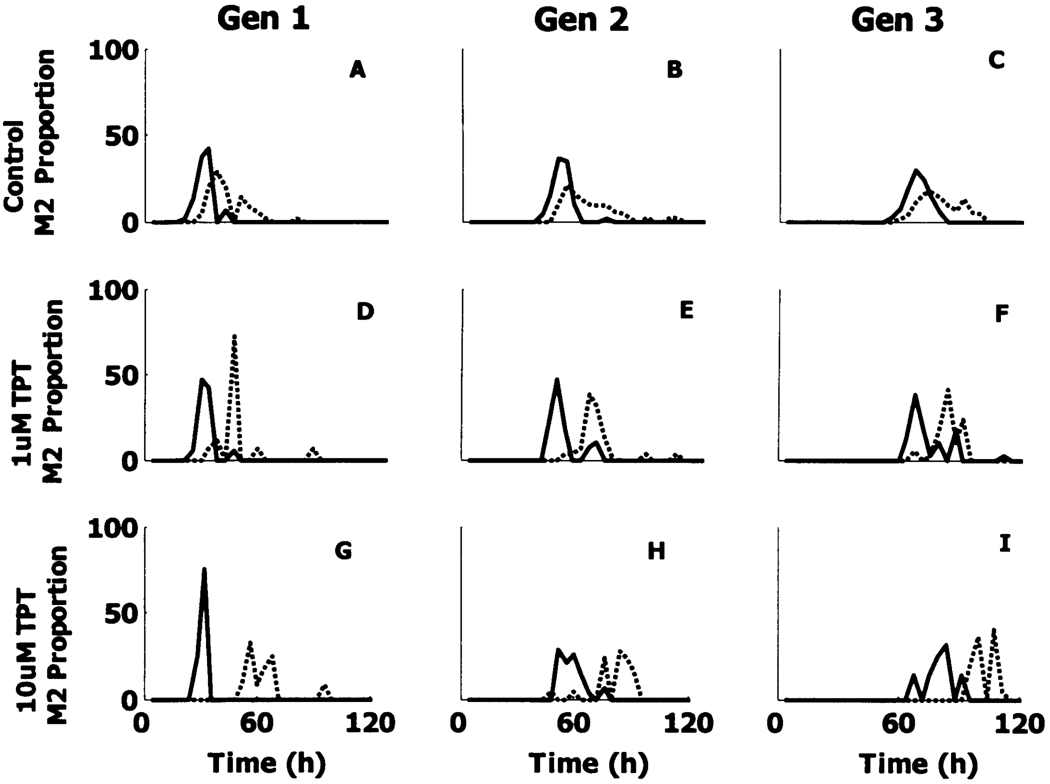


Fig. 4-8 Distribution of mitotic delivery time in a resistant population. Top row of graphs (A-C) shows the mitotic event distribution over three generations in unperturbed (control) condition. Middle row (D-F) graphs after a 1μM TPT condition while the bottom row (G-I) graphs after a 10 μM TPT condition where solid lines represent progeny evolved from G2 origin and dotted line represented non-G2 origin.

The graphs showed two distinctive sub-populations in treated conditions (middle and bottom panel) since the S-phase population was removed providing a clear time gap between the G2 and the non-G2 fractions; with the passage of time (later generations) these fractions started to overlap (see 1 μM TPT condition, Gen 3)., In drug-treated conditions as the S phase specific cell did not deliver to mitosis, therefore no progeny evolved which was represented by the gap within the distribution in progenitor generation, moreover the stress imparted by the drug still had a PD effect on cells delivering to mitosis; with the passage of time this gap became indistinctive as the cells in the later progeny delivering to mitosis in the resistant population (see figure 4-8 F)..However for control conditions (top panel) these sub-populations were not distinguishable at any of the generations.

In order to provide a generalized view of the mitotic distribution simple mean values of both groups in all three generations and conditions were plotted (see Fig. 4-9).

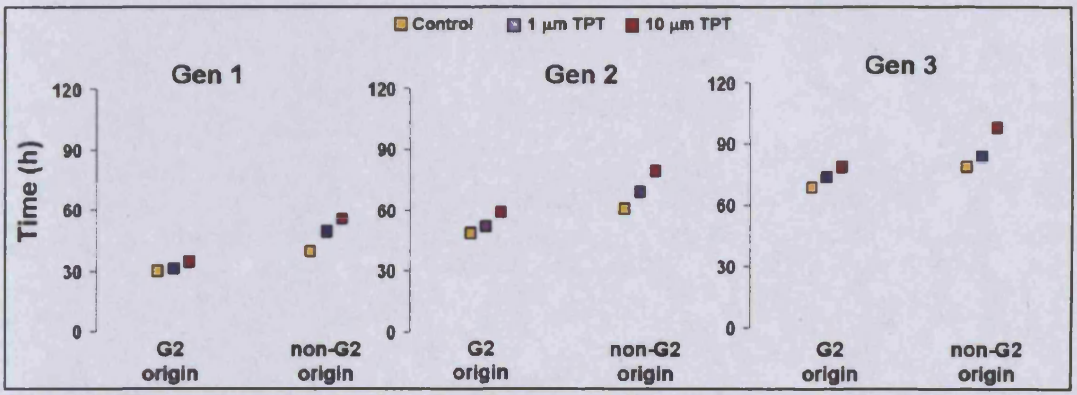


Fig. 4-9 The mean value for the 18 different distribution present in Fig. 4-8 was plotted in three dot plots.

From the plots in Fig. 4-9, it is evident that on average, with the higher dose (10 μM) the progeny delivered to mitosis later compared to lower dose (1 μM), this was more evident with progeny of cells grouped as non-G2 progenitor origin. Interestingly, the cells from G2 origin showed a bigger delay in Gen2 (ie the G2 cell had to go through another cell cycle) after a 10uM TPT treatment. Altogether this implies that if the progenitor cells were positioned in non-G2 phase of the cell cycle when the drug was administered, the resistant progeny evolved will be more affected compared to cells positioned in G2 phase of the cell cycle, hence providing a cell cycle dependent hierarchy for resistance. Again in relation to dose, the average mitotic time for 1 μM condition (violet) gradually became equal to that of control (yellow) and in 3rd generation both for G2 and non-G2 group it was almost identical to control, indicating the drug induced stress disappeared in later generations. However, for a higher dose (10 μM) the apparent delays particularly for the non-G2 origin population still persisted even after three generations.

4.4.1 Statistical analysis using a mixed model analysis

In order to explain these observations and the impending relationships that evolve through the lineages a statistical analysis, called a mixed model analysis, which provided a framework for analyzing data with dependent observations was undertaken in collaboration and input of the expertise of Dr. Valentina Moskvina (analysis was conducted with SPSS 16). The lineage data provide the unique advantage to consign mitotic time (delivery to mitosis) as a dependent observation since this variable could be linked with almost all other encoded parameters of the screen – namely delivery to mitosis time of the mother cell, progenitor cell, sister cell, and other parameters such as those offered by the experimental descriptors such as drug-dose, field of view and well. As in all modelling critical assumptions were made. *The initial delivery to mitosis of the progenitor cells were assumed to be independent of each other and accordingly lineages classified to G2 or non-G2 groups were assumed to be independent.* This is probably not completely a valid assumption since there will be progenitor cells that could belong to the same 'mother' cell, however this designation would have occurred before the start of the timelapse experiment and cannot be deciphered. Thus in this case the possible mother dependency between two progenitor cells is ignored through this assumption. A second assumption was made which removed the dependency between delivery to mitosis time and Field of View (FV) or Well. From a statistical point of view we considered grouping of lineages according to FV or at least to Well, as shown in the Table 4-2, the number of lineages and mitosis were far from equal and to some extent limited. For example in FV1 of Well A1 in control condition of experiment 1, there were 25 encoded lineages of which 23 progenitor cell delivered to mitosis and in successive generations 38, 58 and 94 mitosis were recorded. Within the same experiment 1, if FV10 within Well C2 was considered (10 μM TPT were given) it also had 29 lineages but only 5 progenitor cell delivered to mitosis and in successive generations 6, 6 and 8 mitosis were recorded. To avoid such number discrepancy the level of dependency of lineages were assigned to dose, again even at this level the number of lineages and mitosis were not equal for all groups, i.e. in control condition the number of lineages that deliver to mitosis in successive generation was highest, while for 10 μM TPT it was lowest.

Table 4-2 Number of mitosis with respect to different experimental and generation attributes.

Experiment	Dose	Well	FV	No. of lineages	No. of Mitosis Gen0	No. of Mitosis Gen1	No. of Mitosis Gen2	No. of Mitosis Gen3
EXP 1	**Control**	**A1**	**1**	*25*	*23*	*38*	*58*	*94*
			2	*27*	*25*	*38*	*58*	*92*
			3	*34*	*28*	*49*	*83*	*132*
EXP 2	**Control**	**C2**	**16**	*33*	*30*	*54*	*95*	*155*
			17	*21*	*20*	*30*	*48*	*78*
			18	*22*	*16*	*28*	*38*	*68*
EXP 2	**1 μM**	**B1**	**4**	*19*	*10*	*18*	*30*	*49*
			5	*29*	*10*	*11*	*9*	*13*
			6	*20*	*1*	*3*	*3*	*1*
EXP 2	**1 μM**	**A2**	**10**	*19*	*4*	*8*	*16*	*27*
			11	*19*	*6*	*12*	*16*	*16*
			12	*18*	*7*	*14*	*20*	*19*
EXP 1	**10 μM**	**C2**	**10**	*29*	*5*	*6*	*6*	*8*
			11	*28*	*9*	*12*	*17*	*10*
			12	*28*	*4*	*7*	*8*	*10*
EXP 2	**10 μM**	**A1**	**1**	*20*	*5*	*7*	*7*	*6*
			2	*15*	*6*	*11*	*16*	*21*
			3	*23*	*5*	*7*	*11*	*9*
EXP 2	**10 μM**	**C1**	**7**	*20*	*7*	*10*	*15*	*17*
			8	*24*	*6*	*9*	*11*	*13*
			9	*20*	*7*	*6*	*11*	*15*

Considering the limitations of these assumptions, the dependence of the mitotic delivery time on the progenitor origin and on the dose was investigated for each generation separately up to the 3rd generation. For the first generation the mitotic time was regressed against progenitor origin and dose. For the second and third generation, the dependence due to common mother (division time of sister cells), as well as mother's division time were taken into account as a random effect and as a covariate, respectively.

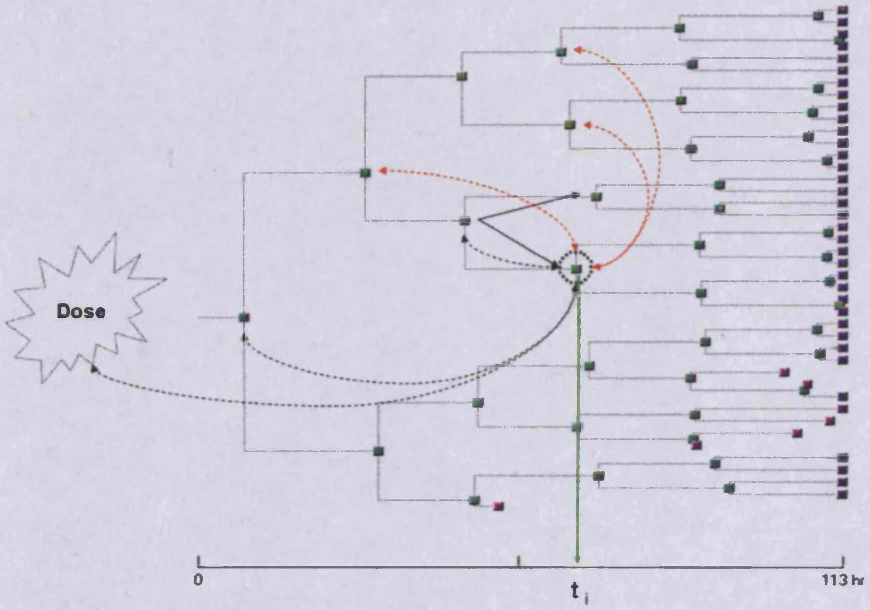


Fig. 4-10 Schematic to show the model analysis dependencies and influences. Delivery to mitosis time (ti, shown by the solid green line) of an exemplar 3rd generation cell (dotted black circle). The black dotted lines indicate the dependency that were taken into consideration (with TPT dose, progenitor origin, mother delivery to mitosis). Solid black line indicates the covariates (mother division time) and indexing two sister cells whose division time were considered as random effect. The dotted red line indicates further relationships and dependency that were not investigated (with grandmother, cousins delivery to mitosis time) for this analysis.

The dependence on grandmother was ignored for this analysis due to the fact that only the 3rd generation cells within this analysis had a grandmother (ie. generation 2 and 1 did not have grandmothers) and moreover these would invoke the consideration of cousin and second cousin delivery to mitosis time etc., which indeed would increase the complexity level of the analysis. Therefore the aim of this study was to first demonstrate the relevance of such a model analysis with lineage data. and secondly to explore simple lineage derived influences to reveal cell cycle dependent action of topotecan.

Table 4-3 Mixed model analysis result

Generation	**Comparisons**	**Progenitor Origin (P values)**	**TPT Dose (P values)**	**Mother's delivery to mitosis time (P values)**
1	Control vs 1μM TPT	$<1x10^{-6}$	0.0002	N/A
2	Control vs 1μM TPT	0.953	0.331	$<1x10^{-6}$
3	Control vs 1μM TPT	0.010	0.553	$<1x10^{-6}$
1	1μM vs 10μM TPT	$<1x10^{-6}$	0.0150	N/A
2	1μM vs 10μM TPT	0.548	0.036	$<1x10^{-6}$
3	1μM vs 10μM TPT	0.039	0.304	$<1x10^{-6}$
1	Control vs 10μM TPT	$<1x10^{-6}$	$<1x10^{-6}$	N/A
2	Control vs 10μM TPT	0.707	0.390	$<1x10^{-6}$
3	Control vs 10μM TPT	0.348	0.381	$<1x10^{-6}$

The analysis undertaken took into account three comparative scenarios, two of dose versus control conditions and the third between the two TPT doses. The outcome from this analysis showed that the 'Progenitor Origin' (G2 or non-G2) and the 'Dose' were significant predictors for the delivery to mitosis in the first generation for all comparisons. Therefore this shows that TPT caused an acute cell cycle dependent-delay however this dependency disappears for further generations, in other words there are no long term consequences originating from the Progenitor Cell classification. However, the 'Mother's Delivery to Mitosis Time' always remained a significant predictor of the daughter mitotic time irrespective of comparisons and generations, therefore there is a local effect persisting from mother-to-daughter. In simple terms these results imply that both progenitor origin and dose have a significant effect on the cell division time of the first generation cells but with passage of time and with the production of later progeny the effect become indistinctive this is further backed up by previous analysis in section (Fig. 4-8 and 4-9). Therefore cell cycle position of the progenitor cell was a determinant factor for future survival of the progeny, i.e. if progenitor cells were positioned in G2 phase of the cell cycle during drug administration the subsequent progeny are less affected than those positioned in non-G2 (i.e. G1) phase.

The findings from this analysis provided for the first time cell-cell and cell cycle dependent relationship outcomes from a simple timelapse sequence, handled via the lineage format. In this case the outcomes were further validated by the considerable a priori knowledge of the action of TPT. The intention would be to use this approach to screen, quantify and visualize these emergent properties in a drug discovery format. The next tool development phase would require more intuitive and interactive features to interrogate the data in such a manner.

4.5 Cell cycle variation between daughter cells – the origin of heterogeneity

Molecular, functional and structural asymmetry for daughter inheritance at cell division is observed in adult stem cells (Chang and Drubin 1996; Guo and Kemphues 1996; Horvitz and Herskowitz 1992; Huang and Raff 1999; Jan and Jan 1998; Kraut et al. 1996; Shapiro and Losick 1997) and perhaps early tumour formation (Wodarz and Gonzalez 2006). In conventional terms asymmetry within a lineage perspective refers to the consequence when one daughter cell delivers to mitosis (i.e. completes cell cycle) while the other daughter dies. This phenomenon may act to impose proliferation advantages or disadvantages acting to widen Darwinian fitness while limiting the divisional potential of a given part (proliferating part) of the lineage. Again within the surviving sub-population, i.e. proliferating part the variation of cell cycling time (IMT) between daughter cells impose another level of asymmetry that leads to heterogeneity in proliferating population. In the previous section it was revealed that the delivery to mitosis of resistant progeny became synchronized with passage of generations, however the extent and origin synchronization between daughter cells was not investigated in that analysis. Considering IMT as a global parameter for the analysis of cell cycle traverse, this section aimed to investigate the degree of IMT asymmetry between daughter cells in relation to generation and different doses of TPT.

The lineage map not only facilitated the ability to visualize and compare asymmetry but provided the means for the identification of daughter cells (hence identified the paired cells) and measure their associated IMTs (paired IMT). LDB was used for this mining purpose and all lineages from the TPT related assays were grouped into three groups according to dose – control , 1 μM TPT and 10 μM TPT. For each group or dose, paired IMT up to 3 generations were identified and sub grouped according to generation. In total 9 dot plots (3 rows for dose and 3 column for generations) were plotted, where the X axis represent the IMT of one daughter cell, termed as South (an arbitrary name for tagging purpose only) while Y axis represents the IMT of the corresponding daughter cell, termed as North.

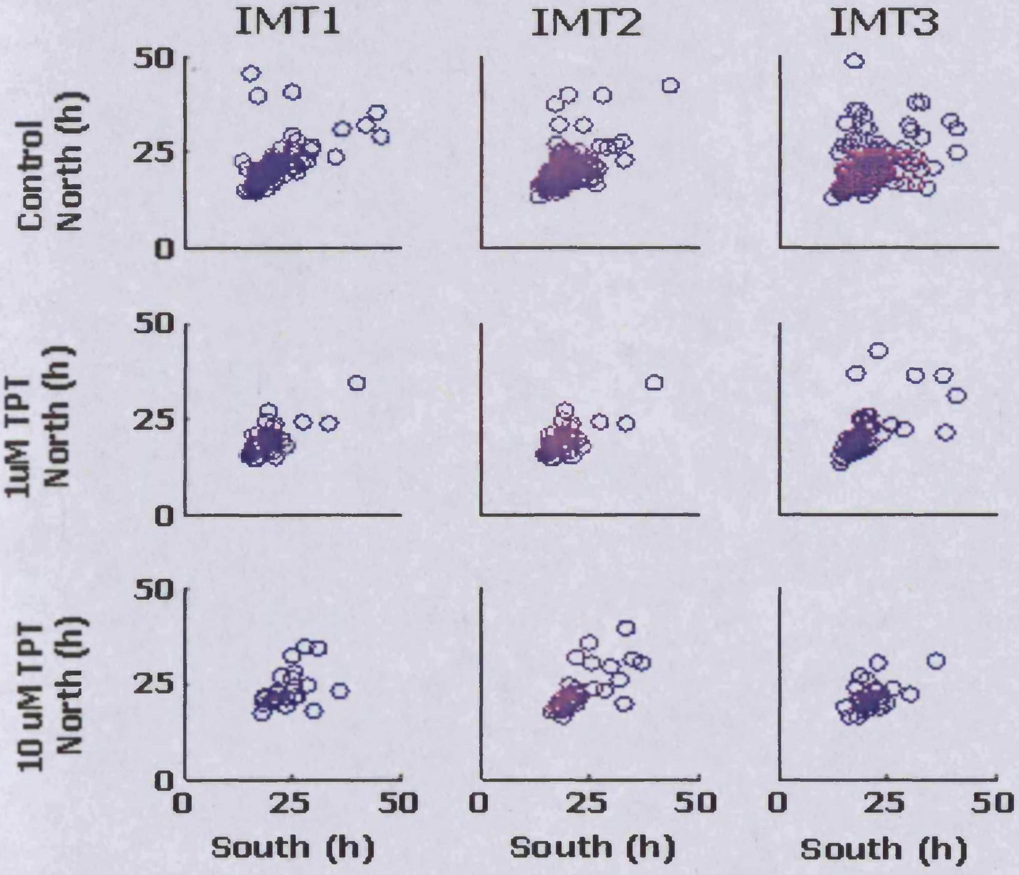


Fig. 4-11 ***Assymetric cell cycle traverse.*** *Scatter plots to visualize asymmetry of IMT between corresponding daughter cells over three generations and in three different experimental conditions.*

Plotting the actual value of IMTs instead of ratio (0-1) served two purposes, primarily this showed the symmetry of the IMTs and secondly showed the dose dependent delay effect when viewed along a column. For example in treated conditions if a point fell outside the diagonal line (of symmetry) and towards the bottom-right of side, it can be explained as – the south daughter becoming more delayed than its counterpart and vice versa (highlighted as red box in figure 4-10). In order to understand this asymmetry through a simple statistical analysis, the paired dataset was subjected to Spearman ranked correlation test (Spearman 1904). Spearman's rank correlation coefficient was used as a measure of linear relationship between two sets of ranked data, in other words it measured how tightly the ranked data clusters around a straight line. Unlike Pearson's correlation coefficient, where it was necessary to assume that both variables have a normal distribution, in Spearman's rank correlation no such assumption was necessary. Spearman's rank correlation coefficient, like all other correlation coefficients, produces a value between -1 and +1. A positive correlation was one in which the ranks of both variables increase together, a negative was reversed. A correlation close to zero means there is no linear relationship between the ranks (Altman 1991).

Table 4-4 Spearman's rank correlation coefficient values obtained from the analysis.

	IMT1	**IMT2**	**IMT3**
Control	0.702672	0.575950	0.444961
1μM TPT	0.458136	0.515487	0.730007
10μM TPT	0.542874	0.816410	0.344143

Only in the control conditions; the correlation gradually decreased with the passage of generations. Two factors were thought to be responsible for this behaviour, firstly with the passage of time the number of data points increased, at 1st generation (IMT1) 101 pairs were analyzed, IMT2 163 pairs and IMT3 275 IMT pairs were considered respectively. Secondly with a higher number of cells in the later generations the crowding effect as described earlier also contributed to this increasing asymmetry of cell cycle traverse. In 1 μM TPT treated condition, at 1st generation correlation was not that evident, however with passage of generations the correlation increased and at 3rd generation (IMT3) showed a high correlation. Same holds for the 10 μM TPT treated condition (except for the third generation) where up to the second generation there was a gradual increase of the correlation. It is important to note here that in both treated conditions, contrary to control conditions the number of IMT pairs decreased, which can be attributed to the increased number of cell death or arrest in later generations. This refers to event based (mitosis vs. death) asymmetry, i.e. predominant source of asymmetry that leads to proliferation advantages as discussed earlier. The sudden decrease of correlation value in the 3rd generation of 10 μM TPT condition could not be explained in a biological context but may be due the small numbers of paired IMTs available for analysis. In this context it is also important to note that the number of IMT pairs analyzed under each group were not nearly equal and as such from statistical point of view was not comparable. Again this limitation is purely due to the small number of encoded lineages which again is directly linked with manpower limitation. Acknowledging these limitations, if the mean correlation value of the three doses were taken, a dose dependent increment of correlation coefficient value (control 0.56; 1 μM 0.57; 10 μM 0.62) was observed which can be explained as follows - for control and low dose the asymmetry of cell division time between daughter cells almost remained the same implying the innate behaviour of the cellular dynamics persists with low dose, but for high dose the cell cycling asymmetry became more correlated, implying that the cell cycle duration of the resistant population in higher dose was more synchronized and increased with time.

4.6 Concluding remarks

This chapter has introduced a new perspective of measuring and visualising cell dynamics that can be exploited in different research contexts and centers around the lineage data format. These 'on-the-fly' parameters (e.g. IMT, IMD, AMT etc.) revealed a time dependent, inheritance based analysis; where cellular behaviour of progeny could be rooted to the causative effect and behaviour of the previous generations, separated by a wide time window. This was important, since in a cellular context the molecular, functional and structural inheritance are the predominant factors that determined population behaviour. Through analysis as discussed in this chapter, it was demonstrated that cell cycle positioning of the progenitor cell was an important determinant for resistant progeny. These informatics-derived results were in agreement with previously found results (Feeney et al. 2003; Marquez et al. 2003) obtained through conventional analysis where overall population behaviour were related with major experimental descriptors to draw the conclusion (Bullen 2008; Lang et al. 2006).

The relationship based measurement provided the foundation for models to predict the future behaviour of cell population. The heatmap view (see figure 4-6 and 4-7) provided a novel approach to visualize the behaviour of a resistant population on successive generations – a PD response in a pedigree format. These behavioural patterns could be utilized as drug signatures in a high content screening environment, where cellular behavioural pattern with unknown drugs could be compared against know drug patterns; which inherently invokes novel algorithms and analysis techniques to be incorporated. More interestingly instead of a drug signature approach, cell behavioural patterns could be examined for genetic manipulation, where cells would be genetically modified (e.g. gene knockout) and the subsequent behavioural pattern would be analyzed to correlate with the gene functions (Rines et al. 2006). This aspiration would open the avenue to integrate '-omic' data with cellular data in a time integrated manner, a perquisite for systems biology (Systems Biology Report 2007).

The lineage data format not only facilitated to quantify the inheritance based behavioural pattern, but also the subpopulation behavioural pattern. In a heterogeneous cell population the overall population behaviour at any given time was the product of subpopulation behaviour (MacArthur et al. 2006), and through the lineage maps it was possible to identify and measure the behaviour of these subpopulations. As the criteria for selecting such subpopulation were not limited by the lineage map, this endeavour can be viewed as an opportunity to differentiate heterogeneous cell population behaviour in a hypothesis driven manner. Heterogeneity of a cell population was underpinned by the asymmetric cellular behaviour (e.g. variation of cell cycle time or cell motility), event outcomes (e.g. mitosis, death etc.) and lineage maps also provided the means to measure and link the extent of asymmetry in terms of cellular behaviour after each cell division point within subpopulation context. These discrete subpopulation behaviour fabricated the emergent properties of the multicellular community. The lineage map along with "microenvironmental" information (Shen et al. 2008) like extracellular matrix substrates, physical forces etc. could be exploited to recapitulate functions observed in native tissues, since exploring this cellular and multicellular form and function remains a fundamental challenge for both biology and tissue engineering (Liu and Chen 2007).

In terms of segmenting the lineage map, it was not limited within vertical and horizontal segmentation only, rather novel and intuitive segmentation and subsequent comparison of 'on-the-fly' parameters can be performed. One such example would be to segment the lineages into two hemi-spheres having the progenitor cell at the mid point (equator) and the effect of perturbing agents can be measured and compared between these two hemispheres. So, the opportunity of exploiting these relationship based data mining remains open-ended. However the prerequisite for all such analysis was the availability of large number of lineages, however manual encoding will always limit the scaling requirement and therefore is a genuine bottleneck. As lineages encoded through either ProgeniTRAK or FluorTRAK (presented in the next chapter) required substantial manual intervention especially for cell tracking and event recognition purpose, the manpower factor becomes the overriding limitation, which in turn constrains the number of lineages encoded. Another limitation that effected the current process of lineage analysis was the consequences of lost events (i.e. when a cell is lost from the field of view), which abruptly truncated part of the lineage, rendering the lineage incomplete. This artefact was different from the dead event that also abruptly truncated a lineage but had a biological significance attached to it. Previous lineage based analysis (Chu et al. 2002; Chu et al. 2004; Endlich et al. 2000; Forrester et al. 2000; Forrester et al. 1999; Prieur-Carrillo et al. 2003) addressed this issue and employed different statistical approaches dealing with missing data points. The next challenge for this project is to identify and incorporate the appropriate statistical approach that would be most pertinent to address this 'lost' cell issue. Work is in progress with statisticians and mathematicians (Prof. Paul Ress's group at University of Wales, Swansea) to undertake a preliminary study in this regard.

In summary the studies presented in this chapter exploited the lineage data format for understanding multi-cellular interactions in a relationship context as outlined in figure 1-2. The future aim would be to incorporate these mining and analytical approaches with ProgeniDB; with an intention of widening the collaborative effort. This would put the lineage format at the centre of the data sharing concept.

Chapter 5: FluorTRAK – molecular tracking onto lineage maps

5.1 Introduction

Cell lineages encoded through ProgeniTRAK have demonstrated to be effective for addressing cellular behaviour and mapping dynamic cellular systems at different levels of complexity as outlined in figure 1-2. The consequences of drug perturbations on the inter mitotic time (IMT) parameter has been explored as a global readout of the cell cycle from generation to generation in the previous chapter to determine the pharmacodynamic (PD) responses to the perturbation caused by the addition of the drug topotecan, moreover cross-sectional data interrogation and merging have revealed cell cycle origins of drug resistance (detail in chapter 6). ProgeniTRAK was originally designed to encode lineages from low resolution image sequences acquired through phase transmission timelapse microscopy, where only major cellular event outcomes could be identified and encoded, however the next phase was to incorporate molecular responses preferably mapped onto the lineage data format at the single cell and progeny level.

Sydney Brenner (Nobel laureate, 2002) has always emphasized the importance of protein localization (Brenner 2003), and stated that the translocation of proteins within different cellular compartments or regions in a dynamic system is necessary for understanding the mechanisms underpinning many biological systems. In this regard he emphasized the urgency of novel reporters and tracking of tagged-proteins in different cell compartments (Yu et al. 2004). Preferably a continuous readout of protein localization at sufficient resolution is required to achieve this goal. Fluorescence microscopy enables the temporal and spatial measurement of single and multiple proteins at a relatively high temporal resolution. This approach has been exploited by many in the context of the cell cycle enabling the tracking of single cell checkpoint transitions in a non-invasive manner even within heterogeneous population. Green fluorescent protein (GFP)-based chimeric probes (Shaner et al. 2008; Tsien 1998) can be constructed to shadow the expression, location and destruction characteristics of endogenous proteins. A good example, appropriate for tracking events in the cell cycle has been based on a GFP shadow reporter (eGFP-cyclin B1) (Thomas and Goodyear 2003), a non-perturbing stealth reporter validated using high content to high throughput detection platforms comprising multi-well high-throughput screen (HTS) imaging, single cell kinetic-tracking and multi-parameter flow cytometry (Thomas 2003; Thomas and Goodyear 2003). eGFP-cyclin B1 tracking provided sub-phase information on cell cycle progression, cell-cycle regulator dynamics in parallel with the lineage morphological landmarks and DNA content analysis. To explore and exploit the continuous molecular level readout for enhancing our understanding on cell cycle studies, a new data format with an incorporated encoding tool was required. The current chapter focuses on the process of developing such a molecular encoding tool – FluorTRAK where the fluorescence levels of the GFP reporter was mapped onto a lineage structure. The later part of the chapter presents ongoing collaborations that exploit the analytical outputs from FluorTRAK and the implications of such data for mathematical modelling.

5.2 Single cell timelapse acquisition – eGFP-cyclin B1 tracking

The encoding of the lineages using FluorTRAK was performed by Janet Fisher and Marie Wiltshire, School of Medicine, Cardiff University.

Fluorescence timelapse experiments over 48 hours were undertaken (for details see Appendix VIII) with the U-2 OS cell line expressing eGFP-cyclin B1. The fluorescent G2M Cell Cycle Phase Marker (GE Healthcare, UK) reporter system was regulated by the control of expression levels and location of eGFP-cyclin B1 as a cell progresses to the later cell cycle stages and negotiates mitotic entry and exit. This was achieved by using the functional components from cyclin B1 to confer switch-like properties to the shadow reporter. Expression was driven by the promoter region, removal via the destruction box (D-box) and translocation from the cytoplasm to the nucleus compartment via the cytoplasmic retention signal (CRS) (see figure 5-1). cyclin B1 expression was tightly regulated and acts as a major control switch suitable for following the transition from S-phase through the G2 phase into mitosis. Importantly since the cyclin box was absent from the reporter it did not interfere with or perturb cell cycle progression (see figure 5-1).

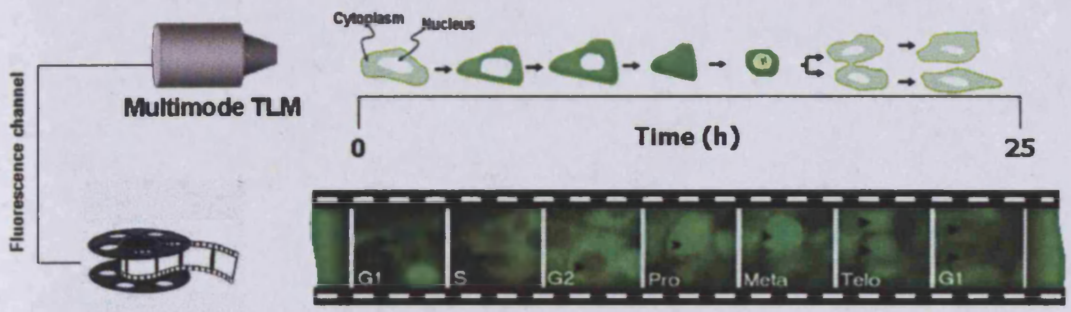


Fig. 5-1 ***Mapping the cell cycle against eGFP-cyclin B1 expression.*** *Top panel, a schematic representation of a cell expressing the eGFP-cyclin B1 reporter as it progressed through the cell cycle to mitosis. Lower panel, real snapshots from a timelapse fluorescence sequence showing eGFP-cyclin B1 expression and hence cell cycle position and progression, the cell indexed by an arrow was tracked through interphase (G1,S and G2) and mitosis (divided into three sub-phases – Prophase, Metaphase and Telophase), before successful cytokinesis into two daughter cells.*

A timelapse microscopy image sequence acquired through a fluorescence channel showed cells traversing the cell cycle and concomitant fluorescence changes as the cell progressed to mitosis from G1, individual cells ramped up eGFP-cyclin B1 expression (became brighter), a translocation event (cytoplasm to nucleus) occurred just before mitosis (see figure 5-1 and for real timelapse sequence see Appendix IX). From an image informatics view an important aspect of the eGFP-cyclin B1 signal readout was whether it was amenable to parameterization and hence potentially incorporated into algorithms for automated analysis and signature identification.

5.3 Encoding cell lineages with a continuous molecular readout

FluorTRAK (the in-house encoding tool) in conjunction with MetaMorph was developed to encode cell lineages to maintain a continuous readout from the single cell while maintaining the lineage relationship similar to ProgeniTRAK. The major difference between these two encoding tools was that for FluorTRAK, data were encoded at each time point (on a frame-by-frame basis), while for ProgeniTRAK data were encoded only when a major cellular event occurred. The encoded single cell eGFP expression data at each time/frame interval when connected, yielded a continuous readout depicting the spatiotemporal properties of the tagged protein associated with all cellular events. The encoding infrastructure as described in figure 2-3 remained the same as that of ProgeniTRAK with the exception that in the 'Encoding' folder, FluorTRAK the encoding tool (also written in PERL see Appendix X) was added and in the 'Experiment Setup' folder the new digital laboratory notebook for fluorescence based experiments – 'Fluorescent_Exp_Setup.xls' was added. New Journal files were also added to the 'Journal' folder that facilitated the selection and upload of the image sequence into MetaMorph with three defined ROIs. Another major difference between FluorTRAK and ProgeniTRAK was that at the single cell level, each cell had two levels of data – i) cell data and ii) frame specific data. At the first level the cell data were similar to the ProgeniTRAK outputs and basically illustrated the event attributes of the lineage map; at the second level, frame-by-frame data illustrated the molecular readout extracted from the three ROIs and encoded the molecular fingerprint at each time point / frame (for lists of attributes see Appendix XI). This modification enabled linking of a continuous molecular readout with event information while maintaining the lineage relationship.

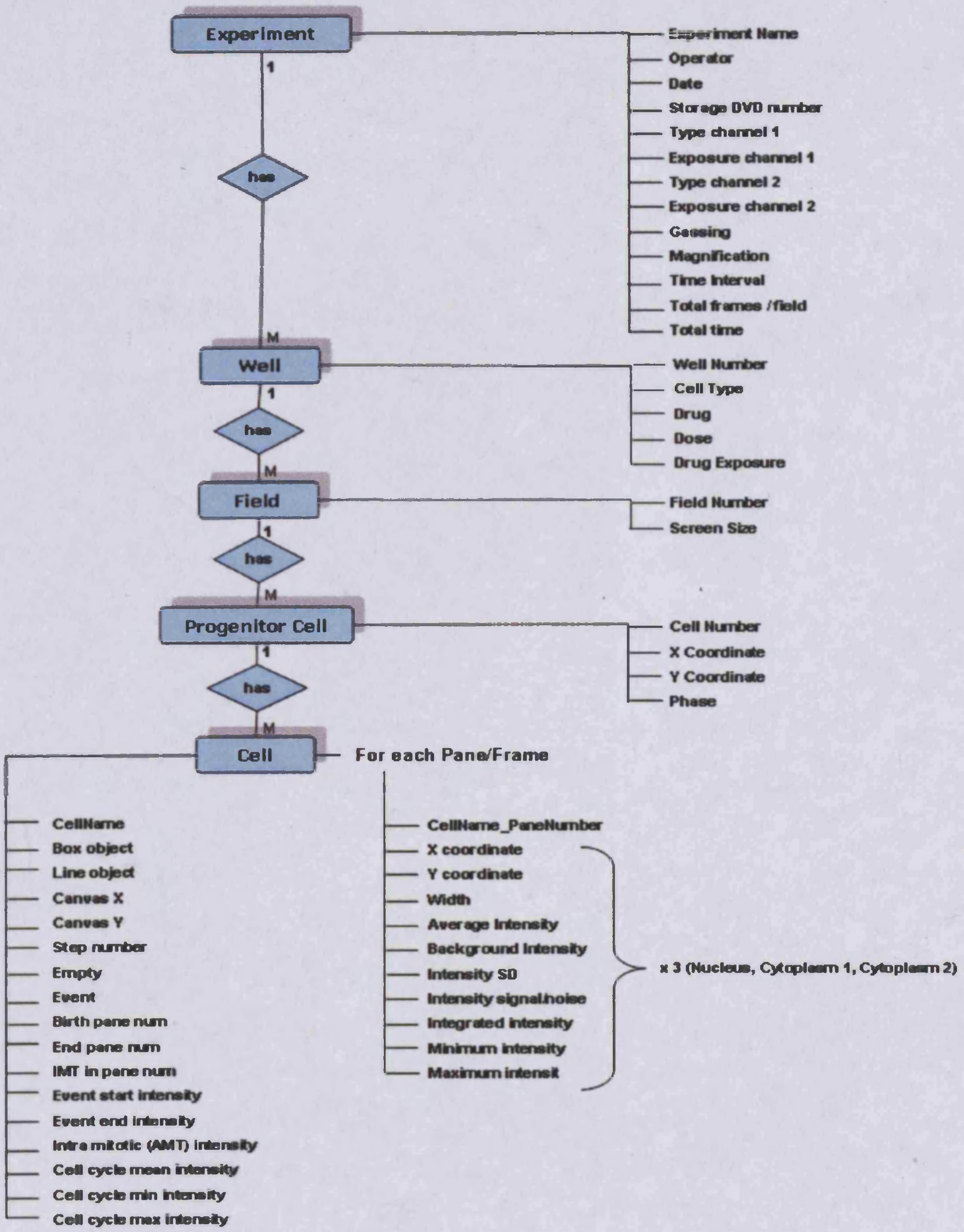


Fig. 5-2 ***Diagrammatic overview of the FluorTRAK data format.*** *At each frame, from 3 ROIs (one placed over nucleus and two in the cytoplasmic regions), 3x10 further data points were collected and appended together to form the molecular fingerprint for each frame.*

The encoding process (for detail see Appendix XII) of FluorTRAK was also different from ProgeniTRAK and indeed was more labour intensive; basically for two reasons - firstly, the FluorTRAK data were encoded manually frame-by-frame for each time interval compared to that of ProgeniTRAK where only major events were recorded. Secondly, 3 separate ROIs were used during the FluorTRAK process. Apart from these two major difference the overall encoding process for FluorTRAK was the same to that of ProgeniTRAK and therefore started with a progenitor cell location; the users would sequentially go through a process of assigning and selecting the experimental attributes based on the screen conditions, including the multi-well details, field descriptors and identifying cell position; additionally a graphical display of the multi-well plate facilitated users in navigating the sequential sieving process. FluorTRAK was designed to dynamically interact with the 'Fluorescent_Exp_Setup.xls' and from the Excel document generated the template and graphical display using PERL. For example, when a particular experimental screen was chosen by the user, FluorTRAK both read and displayed all information regarding each well within the specific screen, the process continued up until the cell level selection. Upon completion of the tagging process, a cell was created on the canvas of FluorTRAK; the corresponding raw image was then retrieved and located in the 'real' image data using the MetaMorph video display window and consequently tracked frame-by-frame. For each cell of interest, in each frame three regions-of-interest (ROIs) were used to extract parameters from the raw image sequence viewed in MetaMorph. The first ROI was always positioned on the nucleus and the other two ROIs were positioned on cytoplasmic regions, usually on the opposing sides of the nucleus of the cell of interest. For each ROI, MetaMorph extracted 10 parameters from the raw image, the parameters included – Frame number, X coordinate, Y coordinate, width of ROI in pixel, average intensity, Intensity standard deviation, intensity signal/noise ratio, integrated intensity, minimum intensity, maximum intensity. Once the ROIs were positioned, the cell of interest was tracked frame by frame starting from the 1st frame. Increment of the frame was automatic when the user presses the 'Log Data' button in the MetaMorph, for any frame if the cell of interest moved considerably from its last frame position, users would be required to reposition the ROIs manually. Additionally when 'Log Data' button was pressed, major and minor events were also logged for that frame. Both event types were displayed in the 'Labelled Logged Data' window of MetaMorph, by default the major event was 'N' and the minor event was 'null' meaning no major and minor event respectively. However with the progression of logging in a frame-by-frame manner, when morphological changes (e.g. rounding up) occurred users would change the minor event to 'start' from its default 'null' label indicating some event had started to occur and finally when the event (mitosis, death) ended users changed the major event label accordingly and minor event label to 'end', indicating the major event had ended. MetaMorph was interfaced with FluorTRAK via the Dynamic Data Exchange (DDE) link and all extracted data along with the associated tags were parsed to FluorTRAK which then according to the major event and the time associated with it drew the lineage within its canvas. During encoding, it was often required to revisit the bifurcation nodes (where one cell divides into two daughter cells) in the image, since only one cell can be tracked at any given time. FluorTRAK has an additional feature through which any bifurcating point of a lineage can be indexed in the raw image sequence viewed under MetaMorph. Once a lineage for a progenitor cell was completely encoded it was saved as a text file. Any lineage can be selected and visualized just by selecting the appropriate lineage text file and indeed editing/revision would also be possible. The editing feature of FluorTRAK provided the opportunity to users to delete any part of the lineage and re-encode if required. These editing rights certainly contributed an added layer of encoding accuracy ensuring quality control of the encoded data. FluorTRAK provided complete flexibility as it can map lineages based on all possible outcomes of a cell division, for example unusual circumstances such as the generation of three or four daughters due to abnormal cytokinesis, or the generation of a polyploidy cell. Like ProgeniTRAK, FluorTRAK also assigned to each cell a unique identifier (name) like B, BS etc. As before this naming approach helped to establish and maintain the relationship between different cells within a lineage.

The semi-automated and user interactive encoding from the raw images was indeed labour intensive, depending upon (i) the size of the lineage, (ii) expertise of the user and (iii) cell density in the image, it took anything between a few minutes to an hour to encode a single lineage. The semi-automated manner of encoding was undeniably the rate limiting step but user intervention ensured the highest precision of the data being encoded. A combination of automated and user-interactive bioinformatics software has been suggested by a recent review as the challenge and opportunity for the next generation of high content screening (Taylor and Giuliano 2005) and recent endeavours (Shen et al. 2006) have demonstrated the potentials of such an approach. The principle problem with eGFP-cyclin B1 tracking using automated approaches is that during the time course the cells loose signal (i.e. in G1) and therefore segmenting a cell becomes a problem. A second fluorescence tag would need to be incorporated to guarantee robust tracking. Once the encoding of a lineage was complete, the lineage dataset was placed into a temporary database where all lineage data were stored in a tab delimited text file format. Like ProgeniTRAK, one lineage constituted a single text file and the name of the text file is the tag assigned to the progenitor cell. The tag or name of a lineage has 23 parameters associated with it, which made it both unique and therefore distinguishable from all other lineages of the 'Lineage folder'. Within the text file each row represented a cell at a particular frame and the 30 columns of data represented the data from all 3 ROIs (10 data points for each ROI). The unique nomenclature assigned to each cell within a lineage enabled access to the data while maintaining intra-lineage relationships, and moreover the nomenclature of the lineage itself facilitated lineage classification based on user defined conditions, e.g. drug, dose. All lineages accumulated as text files were subjected to an automated but rigid quality control check which ensured that all lineages were stored in the correct data structure.

5.4 Results - Continuous cell cycle tracking at the single cell level

Once the timelapse sequences were analyzed and fully encoded, cellular behaviour within these lineages were interrogated and extracted. To acquire a molecular fingerprint for multi-generation tracks, lineages were subjected to a track-wise (horizontal segmentation) interrogation (see chapter 2).

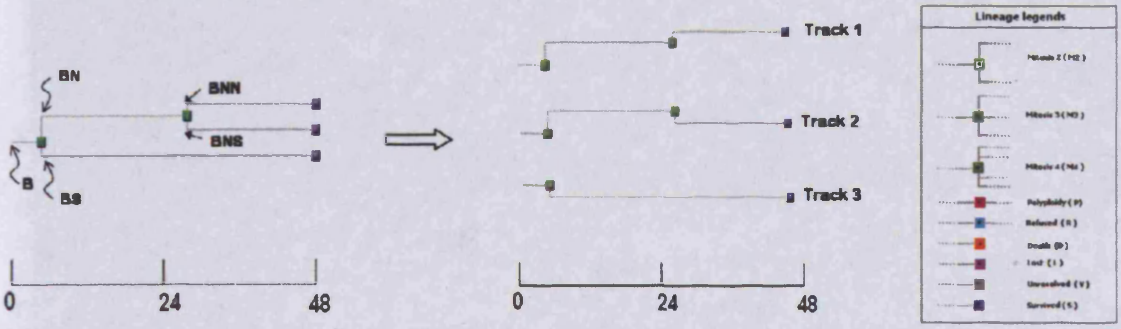


Fig. 5-3 ***Encoded cell lineage.*** *An exemplar lineage encoded from a progenitor cell (B), where the cell divides into two daughter cells (BN and BS) 5 hours after the start of the experimental screen. The north daughter (BN) again divides after 27.66 hours into two daughter cells (BNN and BNS) while the south daughter BS failed to divide within the experimental duration (48 h). Therefore the lineage consisted of three surviving / living cells (BNN,BNS and BS) which at the end of the experiment yielded three tracks labelled as track 1, 2 and 3 respectively.*

The exemplar lineage (Figure 5-3) was analyzed to extract different parameters over time – eGFP-cyclin B1 fluorescence intensity profile, motility and directionality of cellular dynamics.

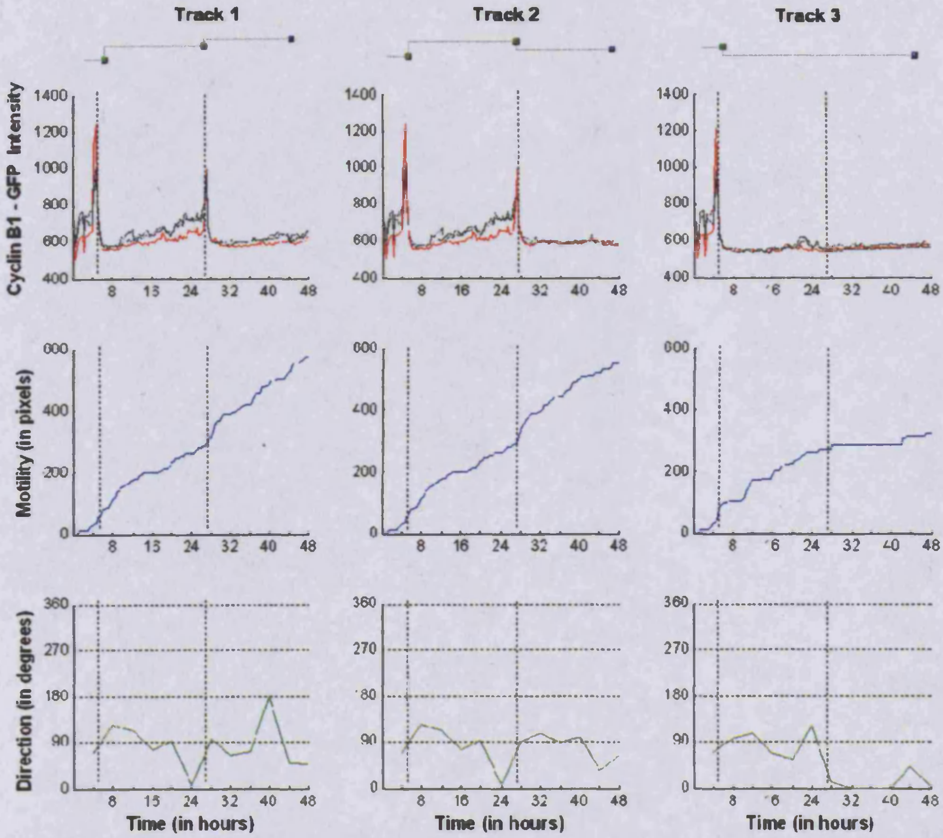


Fig. 5-4 ***Single cell tracking of multi-scalar events:*** *eGFP-cyclin B1 expression (cell cycle progression), distance moved (motility) and vectors (directionality) for each track derived from the lineage shown in figure 5-3. Upper panel depicts the eGFP-cyclin B1 intensity profile along each track, three compartments were tracked - the nucleus (red line) and corresponding cytoplasm (two black lines for two ROIs). Middle panel depicts motility of the same cell (derived from position of nucleus). Motility (in pixels) was defined as the distance travelled by the cell between each consecutive frame (20 minutes interval) and was calculated on-the-fly and was presented in a cumulative manner. Lower panel shows the average angular direction of the cell at every 4 hour time interval. Considering the nuclear position of each cell as the positional point, the tangent angle of each consecutive plane was measured and averaged for each 4 hour interval. NOTE when considering the motility along a particular track, an abrupt increase of the motility corresponding to M phase of cell cycle is observed which is solely attributed to a translation of the dividing cells (i.e. a mechanical artefact).*

5.4.1 Access to molecular fingerprints derived from a cell lineage map

A typical cell lineage over 48 hours illustrated a simple progression of a progenitor cell (B) dividing into two daughter cells and cellular information at two levels: (i) phenotypic behaviour (division and motility) and (ii) molecular readout (eGFP-cyclin B1, hence cell cycle position) at the single cell level (figure 5-4). For the first 5 hours all three tracks were identical, because they were represented by a single cell 'B' which was also termed a progenitor cell. A division occurred at the 5 h time point (node 1) and generated two daughter cells 'BN' and 'BS'. 'BN' again divided into two daughter cells 'BNN' and 'BNS' at 27.6 (node 2). From the point of node 1 to node 2 both track 1 and track 2 were the same and represented by the cell 'BN'. The temporal distance between node 1 and node 2 illustrated a typical IMT of around 22 hours. The reporter tracking of the eGFP-cyclin B1 probe intensity in the cytoplasm started to rise 4 hours after node 1 (in cell cycle terms this would correspond to late-G1), while the fluorescence intensity from the nucleus remained low until node 2 (in cell cycle terms this marks late-G2), the translocation (prophase) occurred just prior to the dramatic change in cell shape from flat to round at node 2. Node 2 marked the mitotic event and the eGFP-cyclin B1 intensity is switched-off and back to basal level both in the cytoplasm and nucleus. The translocation event at node 2 represented a major cellular commitment from G2 to mitosis (M). If the variation of intensity, motility and directionality between two sister cells (BN and BS) from node 1 were compared, no further major cellular event for BS was observed; and the eGFP-cyclin B1 profile remained flat compared to the increasing intensity of BN depicted both in track 1 and track 2. In cell cycle terms this was interpreted as a G1 arrest of BS; by extracting the cell motility it was demonstrated that with this arrest in G1 there was a corresponding halt in cell motility.

Defining cell motility, proliferation and differentiation events is important as they underpin basic biological processes such as growth (Palaniappan et al. 2004b), tissue repair (Farooqui and Fenteany 2005) and metastatic potential (Ronot et al. 2000). Cell motility is thought to be a critical factor for the process of metastasis (Cavanna et al. 2007), the spread of cancer from its place of origin to a secondary site. The molecular mechanism of metastasis is mostly unknown (Jonsson et al. 2006) and recent *in vivo* study has shown that metastatic tumour cells move 4.5 times as frequently as non-metastatic tumour cells in the same tumour microenvironment (Sahai et al. 2005). It has also been suggested that only a subpopulation of cells from a tumour, rather than every cell in the tumour, becomes metastatic (Fidler and Hart 1982). Therefore, it is reasonable to suppose that the increased speed of migration of these subpopulations contributes to the increased metastatic potential. FluorTRAK encoded lineages offers both the measurement of cell motility as well as subpopulation segmentation, which would contribute towards investigating the process of metastasis. The potentiality of such multi-scalar cellular information in a wound healing context was also considered; as links between cell motility and cell cycle check point regulation is of great interest in wound closure and remodelling. Tracking single cell behaviour on induction of a wounding stimulus was a prime requirement that allowed the monitoring of wound repopulation while generating descriptors for directed cellular migration and proliferation. These descriptors can be later interrogated to investigate the interplay between i) cell types, ii) cellular interactions and the surrounding extra-cellular matrix (crucial in directing cellular responses) and iii) repair and remodelling events such as cellular differentiation. Parameterization of single cells in clusters within a complex wound arrangement is considered a challenging bioinformatics undertaking (Bunyak et al. 2006) and has wide applications from cell-based assays to systems biology (Frumkin et al. 2005). It is envisaged that FluorTRAK or a modified version could contribute significantly in this regard as with the present status of FluorTRAK could address most of the challenges set forward in these research areas.

In order to explore the potentiality of the encoded lineage data, a range of data mining and visualization tools were developed using PERL and MATLAB (see http://www.mathworks.com/) and were sent (via email) to the collaborators at Warwick University and Mario Negri Institute for Pharmacological Research, Milan, Italy, for further analysis. These tools were accompanied with encoded lineages (as text files) and installation manual (for detail see Appendix XIII) which facilitated researchers at both sites to install and use these analytical tools. This process demonstrated a step forward towards the aspiration of collaborative research especially at biologist-mathematician interface. Apart from logistical containments (data format, database etc.), the bottleneck for such aspiration is to identify the "common ground" where expertise of both discipline can share ideas (the global objective) without understanding the detailed complexity of individual domain. Data sharing in a meaningful manner is the perquisite for such endeavour, which includes easy accessibility of the data and more importantly translating the biological needs into mathematical terms. Last but not the least translating the results in biological context is the concluding aspect of such collaborative effort.

5.4.2 Profiling the eGFP-cyclin B1 oscillation

This work was conducted in collaboration with Drs Paolo Ubezio and Monica Lupi, Mario Negri Institute for Pharmacological Research, Milan, Italy.

Extracting a molecular fingerprint with phenotypic behaviour at the single cell level was the first step towards understanding linked cellular responses and interactions suitable for mathematical modelling. Encoded lineages were used as experimental data to develop and validate such models, however the first step in this regard was to filter the lineages encoded through FluorTRAK that fulfilled certain criteria: (i) selected lineages consisted of at least one track ending with a cell that is alive and (ii) the track consisted of at least two successive mitosis (nodes), where the first node referred to the mitosis of a progenitor cell. These criteria ensured two aspects– first, effectively a 'resistant population' were selected (defined in our screen as the cohort that survive to the end of 48 hour experiment) and secondly, the eGFP- cyclin B1 profile of a complete cell cycle (IMT) could be mined from the encoded lineage data. The analysis was performed separately for three experimental conditions – control, 1 μM TPT and 10 μM TPT 1 hour bolus treatment. For each condition, irrespective of the generation, complete cell cycles were identified and the associated eGFP- cyclin B1 profile was extracted; 30, 14 and 6 such complete cell cycles were identified for control, 1 μM TPT and 10 μM TPT conditions respectively. Since the experimental duration was 48 h and the mean cell cycle duration for U-2 OS cell line was found to be ~ 20.5 h, the cell cycle generation identified was predominantly from the cell cycle of the first generation (in other words IMT1, see figure 4-5). The analytical tools assisted users to select or deselect relevant lineages as these provided the montage view of the lineages that were selected for certain experimental condition. Once selected the lineage tools automatically wrote the track wise intensity profile to a MS Excel file for further analysis.

For each condition these intensity profiles were then plotted, aligned and normalized (detail in figure 5-5 legend) to depict the average dose dependent eGFP- cyclin B1 profile.

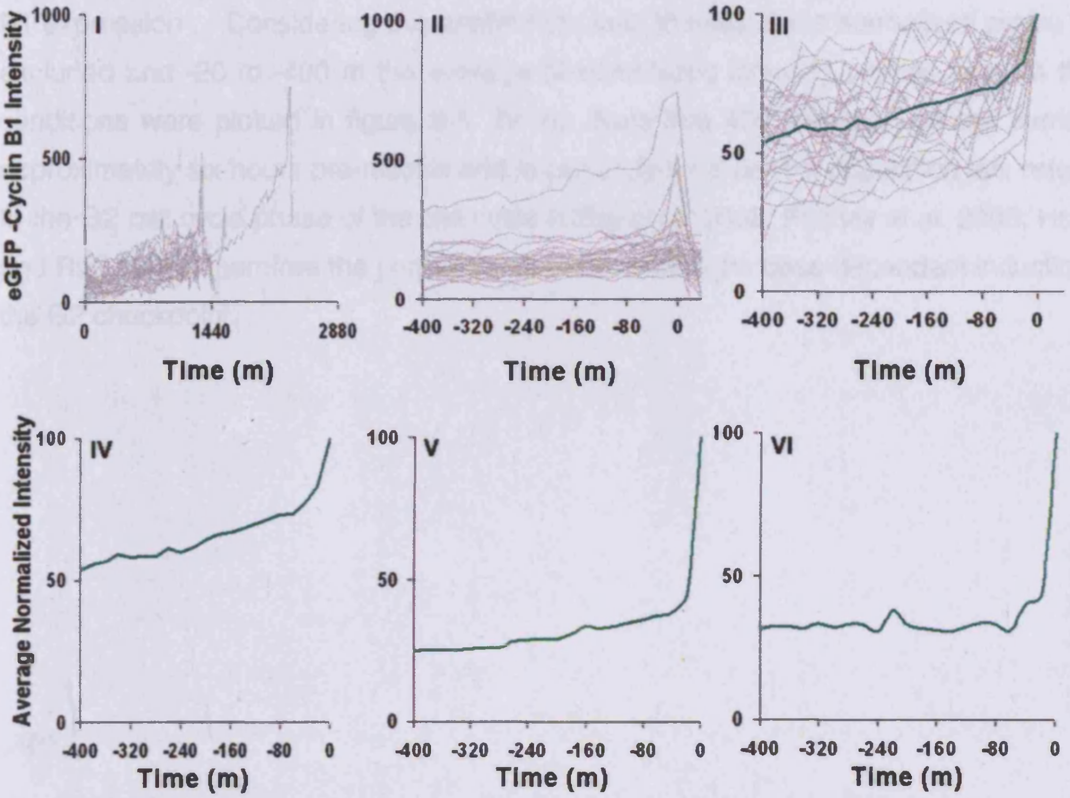


Fig. 5-5 ***Visualizing normalized average cytoplasmic eGFP-cyclin B1 profiles****. Top panel illustrated the plotting, re-aligning and normalization of the eGFP-cyclin B1 profile in control conditions only. 30 intensity profiles were plotted (panel I) with time on the x-axis. For each profile the peak intensity point was indexed as the point of mitosis and set as time 0. Since the duration of cell cycle varies, not all profiles had a peak point at (1230 m ≈ 20.5 h), so all the profiles were re-aligned to their respective time 0 and the last 400 mins up to mitosis of the profile was plotted (panel II). Again the peak intensity value at time 0 varied with individual profile and as such required further normalization. For each profile, setting the peak intensity value as 100, the previous 400 mins values were normalized and were plotted (panel III), the green line represented the average of these normalized profiles. Along the bottom row panel IV, V and VI represented the average normalized eGFP-cyclin B1 profile in control, 1 μM TPT and 10 μM TPT condition respectively using the identical process.*

The eGFP-cyclin B1 intensity profile in the cytoplasm was a good marker to identify and quantify the phase duration for a complete cell cycle. The sudden spike of intensity represented the mitotic event itself and can be attributed to the morphological effect associated with translocation of cyclin B1 into the nucleus from cytoplasm. At this point the cell rounded up and it was difficult to distinguish the cytoplasm from nucleus and thus the fluorescence intensity at this point may not be the true representation of the cytoplasmic intensity, the rounding effect also provided an artefact spike of eGFP-cyclin B1 expression . Considering this artefact the last 20 mins of the normalized profile was excluded and -20 to -400 m the average of normalized intensity profiles from all three conditions were plotted in figure 5-5 IV-VI. Note that 400 minutes was the same as approximately six hours pre-mitosis and importantly for a control population this referred to the G2 cell cycle phase of the cell cycle (Cliby et al. 2002; Feeney et al. 2003; Huang and Raff 1999); therefore the purpose was to determine the dose dependent induction of the G2 checkpoint.

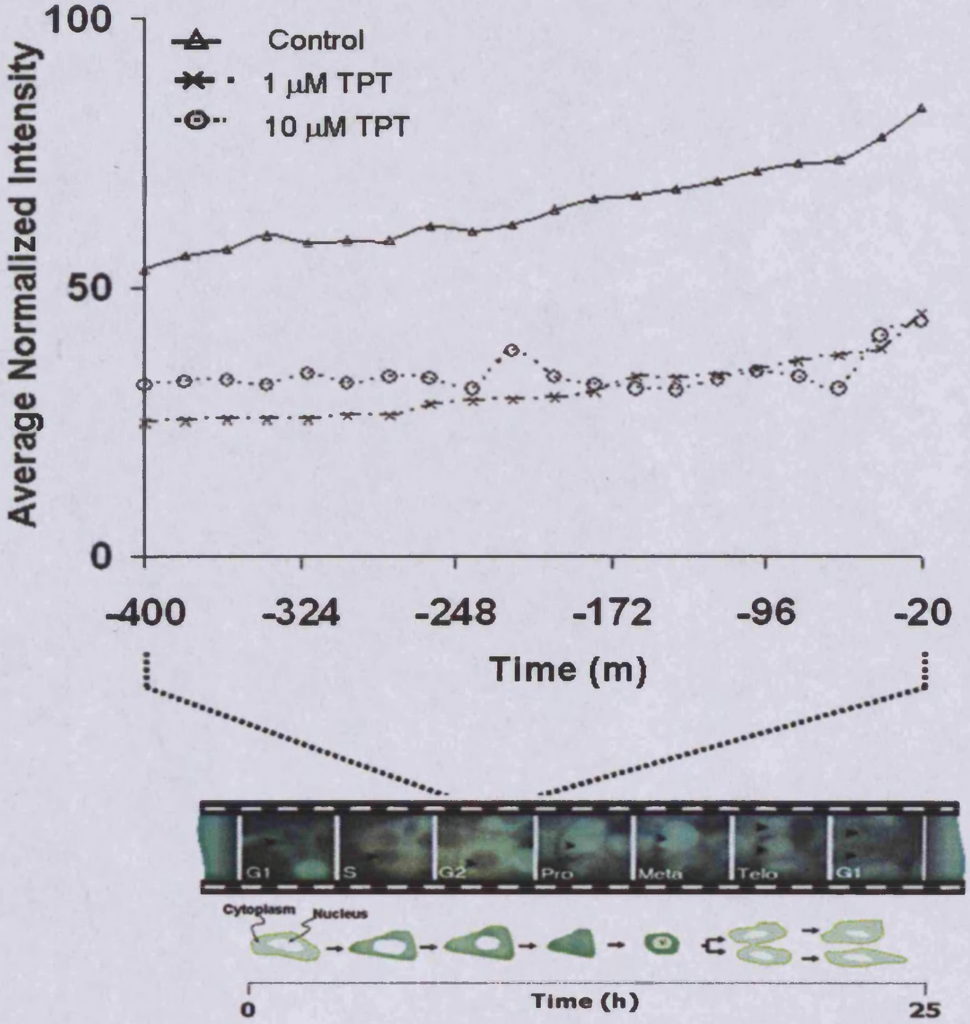


Fig. 5-6 ***Continuous progression through G2****. Normalized average eGFP-cyclin B1 fluorescence intensity through G2 (omitting the final 20 m where translocation occurred) was plotted.*

Both in control and 1 µM TPT conditions eGFP-cyclin B1 levels demonstrated a continuous progression and ramping through to mitosis, indicating the behaviour and engagement however the late G2-checkpoint was shown to be slower in 1 µM TPT treated conditions compared to control. From these two profiles it could be hypothesized that the constant synthesis of cyclin B1 was a good marker to predict the commitment to mitosis. However the flat profile from the higher dose (10 µM) where the 'progression' appeared to be nil, a surge occurred just before the commitment to mitosis this represented a different pattern of G2 progression and mitosis commitment . Interestingly this indicated that cyclin B1 level increase was no longer a good predictor of G2 progression and in some cases not required for commitment to mitosis; i.e. cyclin B1 increase became dissociated from the cell cycle clock (Lindqvist et al. 2004). At the high dose the cells showed a prolonged delay in G2 and the maximum levels of eGFP-cyclin B1 were sustained. G2 arrest of cells due to DNA damage in S phase must avoid entry into mitosis, with the concomitant risks of oncogenic transformation. According to current models, signals elicited by DNA damage prevent mitosis by inhibiting both activation and nuclear import of cyclin B1-Cdk1, the master mitotic regulator (Charrier-Savournin et al. 2004).

Arguably the encoded data had limitations as the data could be considered both qualitatively and quantitatively noisy data. From a qualitative perspective the profiles the image sequence did not provide a consistence output of fluorescence intensity. This was primarily due to the fact that not all cells expressed the same amount of construct and therefore the dynamic range of the reporter varied from cell to cell. From a quantitative view insufficient cell cycle profiles were included particularly at the high dose, which was in-adequate for robust statistical analysis. Despite these limitations, the impact of such enriched data was considered in the context of understanding some of the molecular responses and their complex interplay within the cell level information. Moreover this systematic access to cellular and molecular information can be utilized to develop new cell cycle mathematical models capable of simulating complex cell cycle behaviour.

5.4.3 Building mathematical models based on cell-based molecular readouts

This work was conducted in collaboration with Drs Michael Chappell, Neil Evans and Judith P´erez-Vel´azquez, Department of Engineering, University of Warwick.

The regulatory mechanism of the cell cycle as described in chapter 1 involves multiple proteins (see figure 1-3), many experimental studies have confirmed, that the DNA replication-division cycle in all eukaryotic cells is controlled by a common set of proteins interacting with each other by a common set of rules (Csika´sz-Nagy et al. 2006). Again with the cyclin family of proteins, cyclin B1 regulates the transition from G2 phase to M phase (Kushner et al. 1999). Cyclin B1 binds to cdc25, which then becomes dephosphorylated and relocated to the nucleus (Li et al. 1997), ensuring the transition toward mitosis. Whereas the cdk1 level is typically constant throughout the cell cycle, cyclin B1 expression is cyclic with a minimal expression in G1 phase, an increased level in S phase, and a peak at the G2-M transition (Norbury and Nurse 1992; Pines and Hunter 1994). The continuous molecular readout of cyclin B1 provides the fundamental experimental data for developing mathematical models enabling the simulation of the cell cycle and the focus of such theoretical studies ranges from phase transitions in the cell cycle (Alarcon and Tindall 2007; Novak et al. 1999) to the response of the cell cycle under unique conditions, such as those for cancer cells and the use of anti-cancer agents (Alarcon et al. 2004; Alarcon et al. 2006), including some cyclin B1 - based early models (Goldbeter 1991; Tyson 1991).

Using eGFP – cyclin B1 readouts encoded via FluorTRAK, a mathematical model was developed which described the continuous tracking of cyclin B1 through the cell cycle at the single cell level, including interactions with the cyclin B1 inhibitor, p21 (P´erez-Vel´azquez et al. 2008). The model is an extended version of a transition state cell cycle model by Tyson and Novak (Tyson and Novak 2001) and had been linked with a model accounting for the inhibition dynamics of p21 on cyclin B1 (Pomerening et al. 2005; Pomerening et al. 2003).

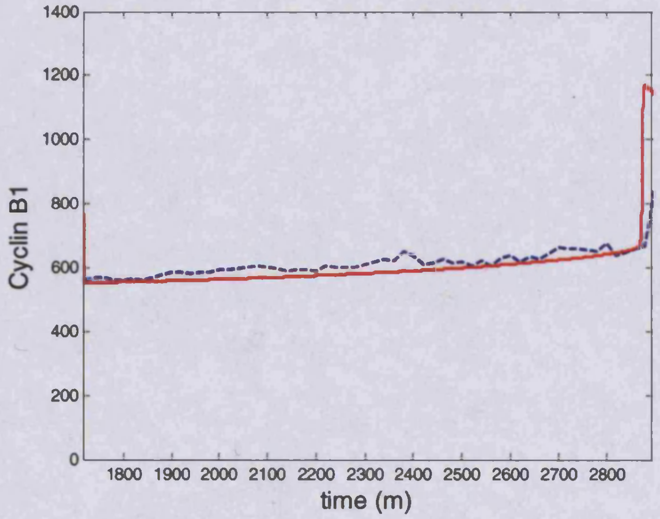


Fig. 5-7 ***Fitting model output with experimental data.*** *Cyclin B1 intensity for one complete cell cycle was plotted, the solid line depicts the simulated model output while the dotted line represented the encoded intensity from FluorTRAK.*

FluorTRAK had multiple implications in the process of model development. Primarily, FluorTRAK provided the cyclin data to the mathematical modeller in a systematic manner, so that modeller had a good orientation and understanding about the experimental datasets. Secondly, the analytical package assisted the modeller to select a set of lineages for training and validating the model. Through exploiting this selection capability another mathematical model was developed that described the response of the growth of single human cells in the absence and presence of the anti-cancer agent TPT (Chappell et al. 2008). The model included a novel coupling of both the micro-pharmacokinetics (PK) of TPT and cell cycle pharmacodynamic (PD) responses to the agent. The model offered the possibility of demonstrating both the dynamic and temporal interactions of active drug delivered to its DNA-associated molecular target and the downstream impact on cell growth and death.

The outputs of the cell cycle mathematical models lie with the ability to undertake parameter estimation, where essentially a fluorescence plot was converted into a molecular [CycB] profiles with 8 new estimated parameters that best described the underlying oscillatory pattern or cell response. The phase plots obtained from such integrated model attempted to predict the length of cell cycle position and variation of which was a marker to identify the amplitude of perturbation. In the previous chapter the amplitude as well as the pattern of IMT in response to different doses of TPT was illustrated (see figure 4-6 & 4-7), however the molecular drivers (parameters in modelling terms) or mechanism that dictate such a pattern could not be revealed. However through model simulation and fitting to experimental data, these occult parameters could be revealed. For example the cell cycle model implemented by P´erez-Vel´azquez et al. (P´erez-Vel´azquez et al. 2008) utilized cyclin B1 data encoded through FluorTRAK but described the role of p21 and through the study of the sensitivity of the parameters, it was possible to identify which (and how) parameters affect important features of the cell cycle, like time between mitotic events, time of first mitotic event and number of mitotic events. The same holds true for the PK-PD model which with its new 8 parameters provided a framework to investigate the role of the signals from damaged DNA to induce the arrest of cell cycle traverse by engaging molecular aspects of the cell cycle machine. It is important to note that the data format of FluorTRAK was generic as it consisted of a multiplex feature which could accommodate up to three simultaneous molecular profiles. This functionality has paved the foundation to analyze and model multiple molecules of interest which from a biological context may be necessary since almost all biological process involve a network of molecules (Wu et al. 2008). In the present context this was not required, but given the ability of present technologies for transfecting cells with multiple tagged molecules and at the same time the design of FluorTRAK has the ability to encode multiple channels simultaneously. The provenance of the new bioinformatics infrastructure provides a platform from which to develop complex mathematical models involving multiple molecules and feedback mechanisms.

5.5 Concluding remarks

FluorTRAK has provided a step change in our ability to encode and access information on a multi-scalar level. Kinetic measurements mapped onto lineage maps have provided an essential route to revealing the critical time windows and informative cells to study the mechanism of action of DNA damaging pharmacological agents. The encoding process encapsulated the critical features of cell-cell heterogeneity, molecular oscillations, phenotypic behaviour and time-dependent events. The multi-level descriptors and parameters attributed to each cell (and at each node), within the resultant cell lineage maps, provided a unique understanding about the high temporal resolution cell cycle phase traverse and checkpoint responses.

Exploiting the encoded lineage data, new mathematical models were achieved. However a number of unmet challenges remain in context of mathematical modelling, predominant of which is the incorporation of stochastic approaches that address the issue of asymmetric division and the inheritance of cell stress . Present models can simulate only one cell cycle under different experimental conditions, but fail to simulate sequential cell cycles addressing the issue of inheritance and stochastic aspects of cell division. This limitation is primarily attributed to the lack of experimental data that incorporates both bifurcation and asymmetry of cell division data. The lineage map provided by FluorTRAK provided a framework to access these features but has conveyed a molecular mapped onto a progeny tree and therefore provided a meaningful structure to the mathematical modellers. These continuous readouts of single or multiple molecules from one or multiple compartments (i.e. from nucleus and cytoplasm) have introduced the opportunity to formulate the next generation of models comprising cellular compartments (spatial) and bifurcation events (division) therefore leading to models that offer prediction *in silico.* From a bioinformatics point of view the next objective is to incorporate interactive features and a "live data sharing" environment (detail discussed in chapter 7) which would enable public access and essentially a community driven research environment.

Chapter 6: Cytometric data linking – the convergence of imaging and flow cytometry derived data to validate and optimize mathematical model

This work was conducted in collaboration with Drs Paolo Ubezio and Monica Lupi, Mario Negri Institute for Pharmacological Research, Milan, Italy.

6.1 Introduction

The previous chapter has established the fundamental concept that the encoded lineage format underpins both the operational (systematic data access) and intellectual (mapping the continuous cell cycle oscillator in a bifurcating system) framework for understanding proliferating cellular systems, leading to enhanced predictive mathematical models (Chappell et al. 2008; P´erez-Vel´azquez et al. 2008). The lineage format challenges the mathematical modellers to contend with heterogeneity, asynchrony and asymmetry in an evolving progeny tree; particularly when dealing with molecular perturbations such as DNA damage and repair. It was discussed that the development of a suitable predictive model cannot be directly coupled to every systems component (e.g. molecular, physiological and network based). Therefore a mathematical model that comprises the 'virtual' cell cycle, including compartmental, spatial and stochastic considerations, fluid dynamics, electrical activities and every gene involved in the regulatory network would not possible, necessary or even desirable (Clyde et al. 2006). Thus a common purpose required from a mathematical model could be defined as the need to deliver the simulation of both the structural and dynamical properties of the biological system. The higher purpose would be to enable a systematic description that reveals the emergent properties which would ordinarily be hidden when viewed from a reductionism point of view (Alfieri et al. 2007). However the principal prerequisite of developing such mathematical models is the availability of appropriate experimental data, generated from wide range of experimental conditions and acquired through different acquisition approaches. We consider diverse cytometric derived data from instrumentation that provided cell-based information at different temporal resolution.

Specifically, data generated by timelapse microscopy with appropriate software like ProgeniTRAK has provided temporal-linked event maps at the single cell level; while flow cytometry data lack this temporal feature but the cross-sectional sampling of a population always gives statistically robust multi-dimensional parameters. Therefore the hypothesis states that the merging of these multi-plexed, multi-dimensional data provides the opportunity to integrate cellular data at the metadata level and also establishes a novel framework for developing a systems approach. Cross platform data convergence has been a long sought bioinformatics challenge and many efforts have been undertaken to converge genomic data with proteomic data in order to undertake computational modelling of regulatory systems in biology, encompassing the regulation of protein complex formation (Reif et al. 2004). However at the cellular level, very few attempts has been undertaken to converge different cytometry-derived data formats and these have been previously reviewed (Systems Biology Report 2007).

As described in previous chapters many studies have verified the utility of timelapse microscopy to determine the intricacies of single cell behaviour (such as heterogeneity, asynchrony and asymmetry) and how different factors may impact on population dynamics (Cervinka et al. 2008; Feeney et al. 2003; Marquez et al. 2003). Transmission phase microscopy offers a probe-less contrast mode providing low resolution but highly informative outputs (e.g. cell shape and cell position) for tracking cell division, cell death, motility providing both temporally and spatially resolved parameters (Farkas et al. 1993). Studies have successfully used the timelapse approach in a screening mode to determine single cell cycle traverse, checkpoint breaching in response to drug perturbations (Marquez et al. 2003; Marquez et al. 2004) and wound closure (Stephens et al. 2004). Furthermore the data have been used to develop mathematical models that simulate cell cycle responses *in silico.* Parallel to our work co-workers (led by P Ubezio, Milan) have previously developed a compartmental cell cycle model for ovarian cancer cells, using flow cytometry data (Montalenti et al. 1998). In essence the model provided an elegant platform for coupling flow cytometry experimental output to a computer simulation, enabling a complete quantitative analysis of the time and dose dependent cell cycle activity and control. This has led to a mathematical model (Lupi et al. 2004) that has reconstructed the traverse of a population of ovarian cells through cell cycle compartments, incorporating delays, and other cell routing such as to cell death and quiescence (G0). Through this model it has been possible to demonstrate that it was not only topotecan (TPT) induced inhibition of DNA synthesis that lead to the cell death, but involved also the induction of G1 and G2-M checkpoints, differential G1 and G2-M block and death, all of which contribute to a specific dose-dependent response. Overall the approach involved deciphering the experimental data (flow cytometric percentages and absolute cell number) by using a mathematical model. Each parameter derived from the model could be viewed as a quantification of the activity of a specific molecular network. The parameters were expressed in terms of probabilities so they were suitable descriptors of inter-cell heterogeneity. Thus in essence the mathematical model provided an intermediate level analysis between the underlying molecular pathways and cellular responses expressed by global parameters like population growth, percentage survival, or flow cytometric derived percentages of cell cycle location and also included cell cycle duration related parameter such as intermitotic time (IMT).

Therefore we undertook a study to determine how to exploit the clear overlap between the timelapse microscopy and the flow cytometry readouts. Model validation is an essential step in the model development process. Most of the mathematical models generated from experimental data are usually cross-validated using a conventional approach where at the preliminary stage, experimental data are split into two subsets – training data and verification data. The former subset is used to estimate the model parameters while the later used to verify the fitness of the model, i.e. validation. However, in this chapter we have taken a novel cross-platform validation approach to validate and constrain a newly constructed cell cycle model for human osteosarcoma cells (U-2 OS cells) (for full details of the U-2 OS model see Appendix XIV). In collaboration with the model originators (Drs Lupi and Ubezio) a detailed flow cytometry-based experimental programme was undertaken to derive the cell cycle model (CCM) output parameters to simulate a U-2 OS population in normal and post-topotecan treatment. The objective of this cross-platform validation was *to investigate how a time series derived from flow cytometry experiments; a time series derived from a Coulter Counter analysis and a time series derived from timelapse microscopy could be linked together to share overlapping descriptors and outputs and therefore could provide a legitimate means for cross-platform integration of data and validation of the model.* The approach is based on the hypothesis that cellular cell cycle properties quantified from different experimental data sources or mathematical model simulations (each termed as platform herein after) depict different aspects of the same cellular system and therefore share common descriptors, which can be exploited to link these platform independent data sources.

6.2 Process of cytometric data integration – the overall schema

The mathematical model developed using flow cytometry data (e.g. bromo-2'-deoxyuridine (BrdUrd) incorporation, see Appendix XIV) remains at the core of this investigation and for simplicity purposes U-2 OS human cell line in unperturbed (control) conditions only was considered. Three types of cytometric data derived from three independent platforms were utilized (i) timelapse microscopy derived Encoded Lineage (EL) data – from ProgeniTRAK, (ii) a time dependent cell count derived from Coulter Counter (CC), (iii) Flow Cytometry (FC) derived output. In conventional cross-validation (Bertuzzi et al. 1988) terms, the cell cycle model (CCM) would have been validated against FC data (the iii platform) alone, where the initial experimental FC data would had been divided into two subsets, one for model development and other for validation. However in this investigation, the model was not only compared or validated against the later subset (reffered as FC data) but also cross-interogated against other two types of data derived from EL and CC, to determine the similarity of two time-series signals (cross-correlation of sorts) and further applied to optimize and validate the CCM. The approach of cross-platform validation as well as integration offers three benefits. First the approach ensures robustness as well as experimental data independence, i.e. the model developed from one source of experimental data can be validated against data generated from different experimental source. Second, integration of cross platform cytometric data provides a means for different data sources to differentially constrain and define the boundaries of mathematical models. Finally parameter linking of cross-sectional molecular (i.e. snapshot BrdU incorporation) data with continuous data (i.e. lineage maps) alleviates the requirement and indeed the data burden for increasing amount of continuous outputs. In fact as described before with current technology, it is not possible to be able to track every cellular and molecular event simultaneously using imaging, and therefore *the work conducted in this chapter has the higher purpose of addressing how cross-sectional data can inform on continuous data outputs and vice versa.* The process of cytometric data integration which at the same time provided cross-platform validation for the cell cycle model, was achieved in two distinct steps.

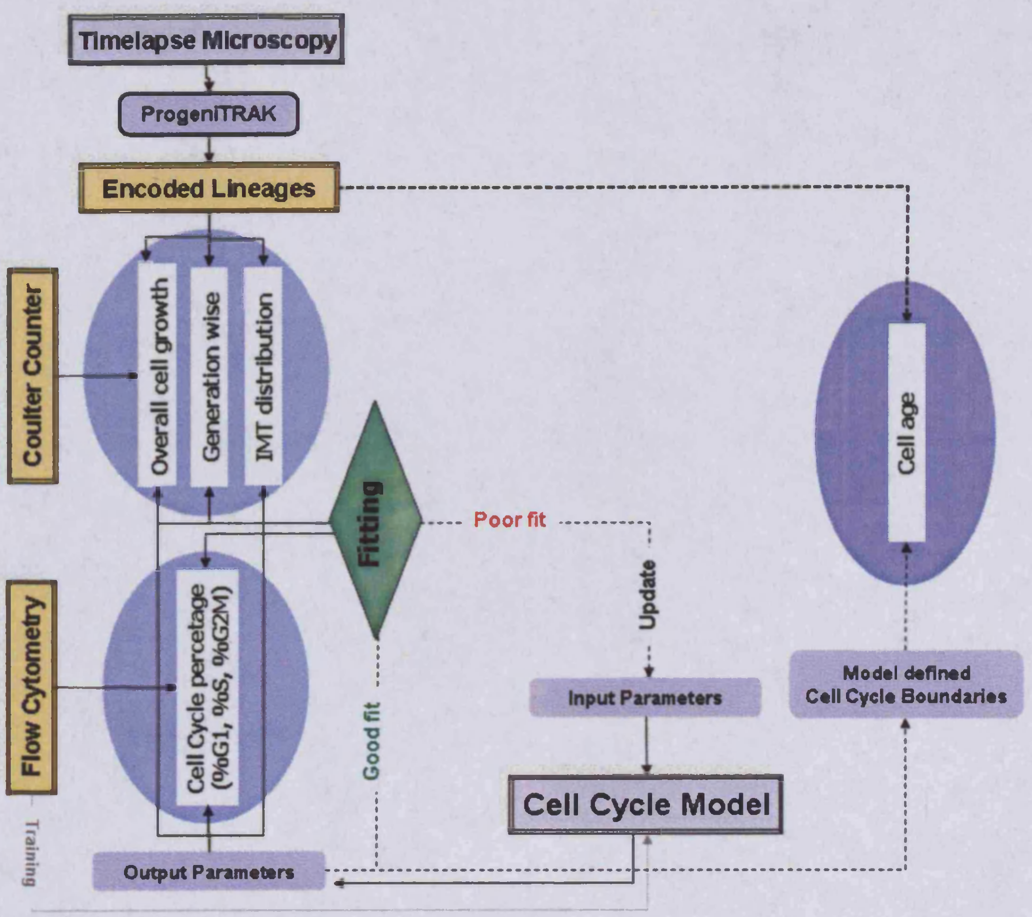


Fig. 6-1 ***Cytometric data integration and model optimization*** *Comparison and integration of cytometric data generated from timelapse, flow and a bespoke mathematical model. Two different areas for integration are depicted (blue ovals) and within each oval different aspects of comparison and integration are listed. The first step, represented on the left, involves the comparison of the data generated from all three platforms (yellow coloured boxes) to optimize the model. The second step, represented by feeding into the linked area (blue oval on the right), involves exploitation of the optimized cell phase boundary values using these generated parameters as markers to assign cell age to cells within the encoded lineages.*

The first step depicted by two ovals (blue) (figure 6-1) involved model optimization as well as cross platform cytometric data integration. The CCM had a number of input parameters (see figure 6-2) which was updated by changing the values for the parameters as well as incorporating new parameters. With updated input parameters the model generated a simulation which in turn spawned a set of output parameters (see figure 6-2). These output parameters were then subjected to comparison with experimental data generated from the three cytometric platforms: FC, CC and EL for comparison as well as optimization. However not all output parameters were compared against all cytometric platforms, rather a partly overlapping approach was adopted. To exemplify, the percentages of cells allocated to different phases of the cell cycle could be best measured from FC data (based on DNA content), but could not be measured from either EL or CC data. Accordingly percentages of cell phases generated from the model output could only be compared against the FC data (as depicted by the bottom left oval in figure 6-1). Again generation wise cell growth and IMT distribution parameters output could only be compared against EL data (as depicted by the upper left oval in figure 6-1) and overall cell growth parameter could be best compared against both EL and CC data (as depicted by the upper left oval).

The purpose of this step was to constrain and optimise the CCM by comparing model output parameters with cytometric data derived from the other three cytometric platforms. The simulation process works as a trial and error procedure where the iterative process continues until a set of parameters was achieved which yielded the 'best fit' with all three platforms derived data (Figure 6-1). Here it is important to remember that both EL and CC data were not only from different platforms but also from distinct experiments so achieving a 'good fit' of the model output against these two platforms provided independent data and can be termed as a process of cross-platform validation. While achieving a 'good fit' against FC data can be termed a process of cross-validation. Both these validations were performed simultaneously and the collective process implies optimization of the model; once optimization was achieved, another level of input parameters such as cell phase boundary(s) (CPB(s)) (which defines the cell cycle phase durations) were then utilized in the next step of the integration process. CPB is an important and a highly sought parameter of cell cycle compartmentalisation that at present can only be quantified from an end point assay system e.g. flow cytometry. However, it is important to quantify CPB in a time dependent fashion and also the amplitude of variability in relation to drug treatment as this retrospectively provides details of drug action.

The second step (depicted by the right-hand side in figure 6-1) in the integration process involves exploitation of the optimized CPB value by using it as a marker to identify and assign the cell cycle phase and the age for cells within EL data (lineage data generated from ProgeniTRAK in this case). Once the age of the cells to EL data were assigned the resultant distribution at each phase of the cell cycle was compared with that measured from model derived data.

6.3 Cell cycle model simulation and parameter estimation for U-2 OS human osteosarcoma cells

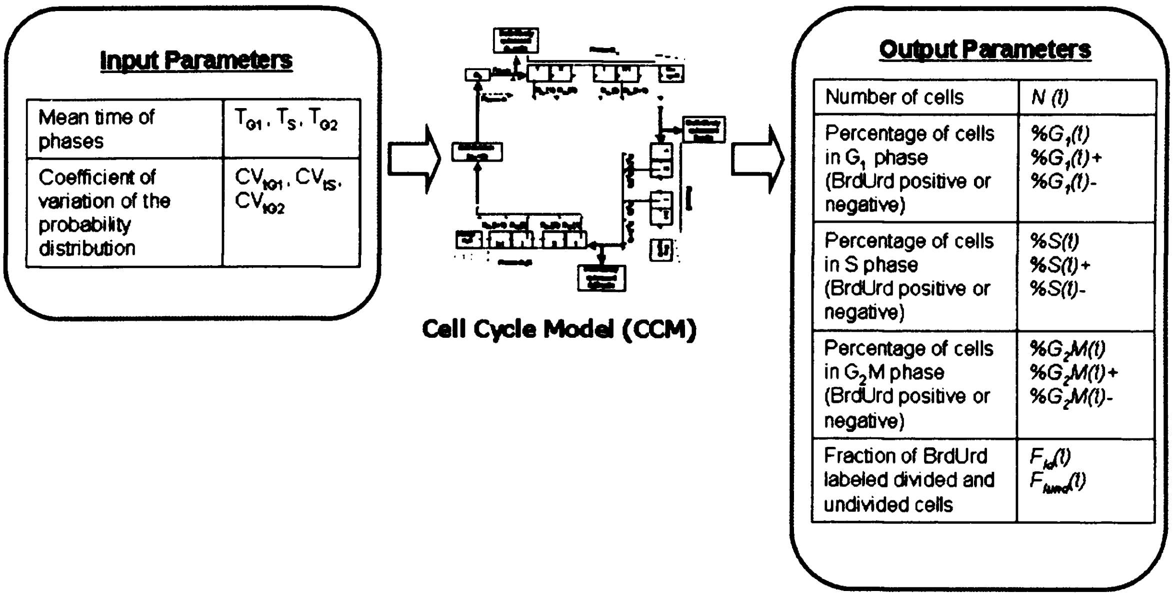
An extensive flow cytometry experimental study was undertaken to establish the compartmental cell cycle model for U-2 OS cells (unpublished, see Appendix XIV for details). Briefly, the cell cycle model was developed on FC data represented as DNA content and bromo-2'-deoxyuridine (BrdUrd) incorporation (Maszewska et al. 2002); BrdUrd replaces thymidine during DNA synthesis, catching cells that are in S-phase only during the pulse. BrdUrd pulse-chase analysis (Higashikubo et al. 1996) was used to assist as a cell cycle marker to reveal the percentages of cells at different cell cycle phases along with many other parameters; and to determine the fate of S-phase cells within a highly sampled time series (0-72 hours) (Lupi et al. 2004). In addition for the first time, data derived from timelapse microscopy were considered together with cell count outputs and flow cytometric (FC) data processed via the U-2 OS cell cycle model (CCM). Overall we combined time-course measures with different experimental techniques and with the aid of a compartmental model simulating cell cycle progression. This mixed experimental-simulation approach enabled us to decode the dynamics for the unperturbed growth of U-2 OS cells (i.e. control conditions) (see figure 6.2). Initial input parameters were estimated. The determination of these parameters can be achieved with progressive levels of complexity:

1. **Constant phase durations** (T_{G1}, T_S and T_{G2M}). In this case it was assumed that there was no inter-cellular variation in the duration of phases and the cells proliferate without perturbations (for details see section 6.3.1).

2. **Variable phase durations**. Input parameters were the mean transit times in each of the cell cycle phases (T_{G1}, T_S and T_{G2M}) and the inter-cellular spread of G1, S and G2M transit times, measured by the respective coefficients of variation CV_{G1}, CV_S and CV_{G2M} (for details see section 6.3.2).

3. **Variable phase durations and cell cycle perturbations**. Additional parameters associated with cell cycle perturbations with an underlying biological significance

were considered in this case (for details see section 6.3.3).



Mean time of phases	T_{G1}, T_S, T_{G2}
Coefficient of variation of the probability distribution	CV_{tG1}, CV_{tS}, CV_{tG2}

Number of cells	$N\,(t)$
Percentage of cells in G_1 phase (BrdUrd positive or negative)	$\%G_1(t)$ $\%G_1(t)+$ $\%G_1(t)-$
Percentage of cells in S phase (BrdUrd positive or negative)	$\%S(t)$ $\%S(t)+$ $\%S(t)-$
Percentage of cells in G_2M phase (BrdUrd positive or negative)	$\%G_2M(t)$ $\%G_2M(t)+$ $\%G_2M(t)-$
Fraction of BrdUrd labeled divided and undivided cells	$F_{ld}(t)$ $F_{lund}(t)$

Fig. 6-2 ***Input and output parameters underpinning the cell cycle model. For detailed CCM see the appendix XIV.***

6.3.1 Constant cell cycle phase duration

Assuming constant cell cycle phase duration, Steel's formulae (Steel 1977) can be applied to determine the fractional duration of cell cycle phases from %G1, %S and %G2M obtained from flow cytometric analysis during exponential growth:

$$\frac{T_{G2}}{T_C} = \frac{1}{\ln 2}\ln\left[1 + \frac{\%G_2}{100}\right]$$

$$\frac{T_S}{T_C} = \frac{1}{\ln 2}\ln\left[1 + \frac{(\%S + \%G_2)}{100}\right] - \frac{T_{G2}}{T_C}$$

$$\frac{T_{G1}}{T_C} = 1 - \frac{T_S}{T_C} - \frac{T_{G2}}{T_C} \qquad \text{Eq. 8}$$

From these formulae the fraction of T_C spent in each inter mitotic phase starting from the knowledge of the percentage of occupation was obtained. Using cell count data, the doubling time and hence T_C can be independently calculated. Thus T_{G1}, T_S and T_{G2M} were estimated as T_{G1} = 5.6 h; T_S = 12.3 h and T_{G2M} = 7.2 h respectively and used as input parameters in CCM. The output plots are shown (figure 6-3) together with data obtained from other experimental platforms.

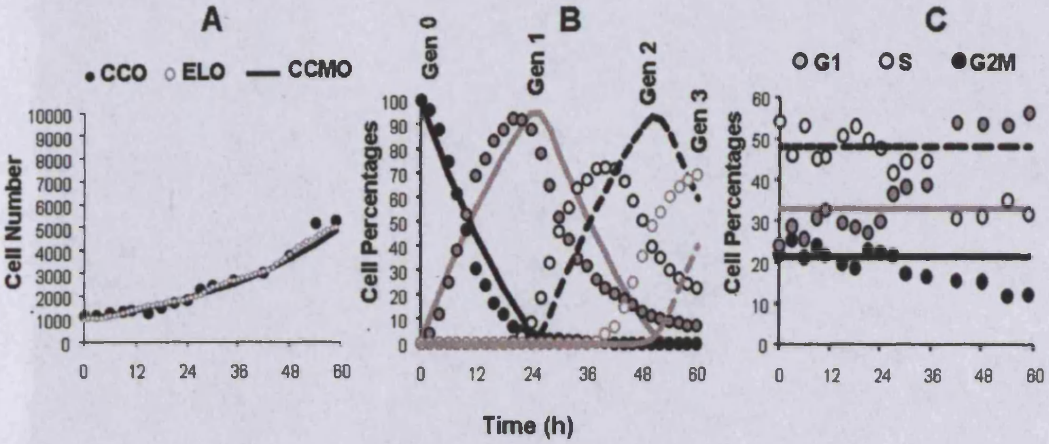


Fig. 6-3 ***Panel A:*** *Overall cell growth kinetics derived from multiple cytometric platforms. Normalized experimental cell counts from the Coulter Counter Output (CCO) [filled circles]. Normalized cell counts calculated from the Encoded Lineage output (ELO) [grey open circles]. Cell cycle model outputs (CCMO) [solid line].* ***Panel B:*** *Comparison between cell generation distributions as obtained from ELO (symbols) and CCMO (lines).* ***Panel C:*** *Experimental FC data (symbols) were compared to CCMO (lines), obtained supposing that the cell population is asynchronously growing with a constant cell phase duration. As shown by the percentages of cells in G1, S and G2M at 0 h this assumption provides an adequate simulation of initial cell distribution, but the presence of additional effects needed to be taken into account in order to fit cell cycle percentages.*

Population growth was independently measured from CCO and ELO. In simple terms cell growth was defined as the number of cells at any given time N_t. The CCO sampled every 3 hours provided the only direct experimental means for counting cells in suspension and was considered as the crucial output for integration as it gave an absolute measurement for U-2 OS cell number per ml at a given time point derived from a whole culture. In the case of ELO, analysis was performed at a time interval (tv) of 30 minutes and at each time Eq. 2 was used again to calculate the value of N_t, where N is the number of live cells at time t.

$$N_t = N_{t-tv} + \sum Mitosis_t - \sum (Death + Lost)_t \qquad \text{Eq. 2}$$

By considering that the rate of cell division or mitotic delivery actually drove the value of N_t and therefore contributed to a net increase, while the rates of cell death and loss decreased this value, even though, in controls, the contribution from cell death was considered negligible and the number of cells lost from the field of view can be replaced at any time by those coming in the same field of view. So while counting the N_t in case of ELO, the negative part of the equation (i.e. cell death and lost) was ignored. As shown in figure 6.3 A, a good correlation was found between ELO, CCO and CCM data, but this was not sufficient to consider the values obtained from Steel's formulae a good estimation of cell cycle phase durations, i.e. T_{G1}, T_S and T_{G2M}

Analyzing cell growth generation-by-generation provided another example of data convergence that was derived from CCM and EL only. It was only from the cell cycle model and the encoded lineages where generation-to-generation information was attributed to individual cell under consideration and thus overall population growth was segmented in terms of generation, both the Coulter Counter and experimental flow cytometry output data did not contain this resolution of information and as such were not included at this stage of comparison. The generation-based information provided the opportunity to explore and compare population growth at different generations in relation to time. The concept was that to estimate cell number at any one time and to assign this to an appropriate generation the following was true - when a cell proceeded through cell division, two new cells were added to the new generation and one cell was subtracted from the current generation (i.e. removing the cell completely from the analysis). Both the cell cycle model output and the encoded lineage output provided the number of cells delivered to mitosis at a designated time point together with associated generation information. Therefore by using the simple principle of counting and transferring a cell count from one generation to the next as they deliver to mitosis (M2), the time and generation dependent fraction of cells at each stage (percentage (%) of cells) was calculated and plotted (figure 6.3 B). It was evident that by taking this view of the data based on generation, the overall oscillatory nature of both the CCM and EL data were quite similar. The comparison of these data revealed a good agreement between the two data sources at generation 0, but started to subsequently diverge from generation 1 onwards, the simulation output shifted in time such that the CCMO always followed the ELO.

Despite the poor correlation of simulated data with experimental data it was clear that even at this simple level of convergence, the timelapse lineage data or indeed the Coulter Counter derived data could be used to constrain the mathematical model parameters. Therefore at this basic data resolution the result implied a depth at which data convergence of different cytometric data source could be achieved. A similar inference could be made from the figure 6.3 C, in this case CCM was able to reproduce the trend of cell cycle percentages as obtained by flow cytometry experimental data analysis of DNA content and bi-parametric DNA/BrdUrd data (see Appendix XIV) for up to 24 hours.

The distribution of inter mitotic time (IMT) provided another important data output that was derived from encoded lineage outputs and was compared with CCM. From the cell cycle model, at a single cell level the IMT value was calculated from summing the individual duration of the cell phases. Therefore cell cycle traverse time T_C for each individual cell was calculated as follows:

$$T_C = T_{G1} + T_S + T_{G2M} \qquad \text{Eq. 9}$$

The cell cycle traverse time T_C was defined as IMT (see chapter 4) and again the encoded lineage data provided access to this calculation. At a single cell level it was calculated as the difference between the first appearance time (A_t) and dividing time (D_t).

$$IMT = D_t - A_t \qquad \text{Eq. 10}$$

Individual T_C and IMT calculated from each data source were binned to their nearest corresponding hour bin and the distribution was presented as a histogram normalized for 5000 cells using the following formula:

$$N_{norm}(t) = N_{(t)} \times 5000 / N \qquad \text{Eq. 11}$$

where N_{norm} is the normalized number of cells at the time t, N_t is the number of cells at time t as obtained from ELO or CCM and N is the total number of cells present during the whole experimental duration.

The model at this stage was not considering heterogeneity in $T_{C.}$ The comparison of the experimental data with the IMT distribution as obtained from CCM (figure 6-4) followed the assumption of constant cell cycle phase duration and revealed the requirement to consider a normal distribution of T_C in order to fit that data derived from EL.

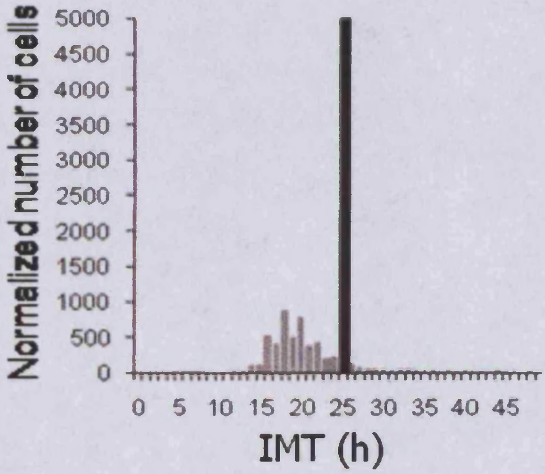


Fig. 6-4 ***Distribution of IMT****. Gray bar graph displaying the distribution of IMT extracted from EL data while the black bar represents the CCM generated output.*

6.3.2 Inter-cell variability in cell cycle phase duration

The mean transit times of each cell cycle phase T_{G1}, T_S and T_{G2M} and the inter-cellular spread of G1, S and G2M transit times, measured by the respective coefficients of variation CV_{G1}, CV_S and CV_{G2M} were altered in CCM in order to fit CCO, ELO and FC data, but direct reproduction of IMT data was not feasible with the program. Thus the same frequency distributions of phase durations adopted in the model were used to generate T_{G1}, T_s and T_{G2M} values with a Monte Carlo routine. A set of $Tc = T_{G1} + T_s + T_{G2M}$ values were obtained in this way, the generated frequency distribution was compared with lineage derived IMT distribution. Monte Carlo simulations demonstrated that a shorter Tc (20.1 h) and a CV around 30% in all phases (Table 6-1) provided a better fit between the IMT data and CCM derived T_c data, at least for Tc that are not too long, as shown in the figure 6-5.

Table 6-1 Cell cycle phase *duration with respective coefficient of variation*

Phase Duration	Coefficients of variation
T_{G1} = 3.9h	CV_{G1} = 30%
T_S = 10.3h	CV_S = 30%
T_{G2M} = 5.9h	CV_{G2M} = 30%

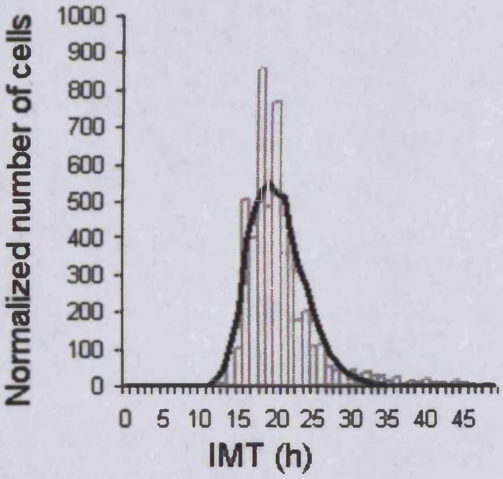


Fig. 6-5 ***Fitting of IMT distribution***. *Solid line represent CCM data output, grey bar graph represents EL data.*

Applying the new input values obtained in Table 6-1 into the CCM, a similar process was undertaken as before (figure 6-3) to reproduce output plots and was presented in figure 6-6. When considering cell number increase over time figure 6-3 A and 6-6 A, the fit was poor (worse than before), however the fit of cell generation distributions was much improved (figure 6-6 B) and that of cell cycle percentages (figure 6-6C) was almost unchanged. These results imply that the consideration of CV was required but it is not sufficient and as such additional variable effects (perturbations) were required to fit well across all experimental data.

Even in cell populations growing without the addition of any perturbing agent it is possible to observe the presence of a synchronized sub-population that moves slowly through the cell cycle. These differential sub-populations originate from many potential sources and can be ascribed to cell manipulation, especially when observed at short times after BrdUrd labelling, as BrdUrd labelling and washout may induce temporarily perturbation in cell growth (such as a slight accumulation of cells in a particular cell cycle phase) this effect was consider as "cell manipulation". Again the differentiation of sub-population may also caused by clonal variation within the population and confluency effects when observed after a few days of culture.

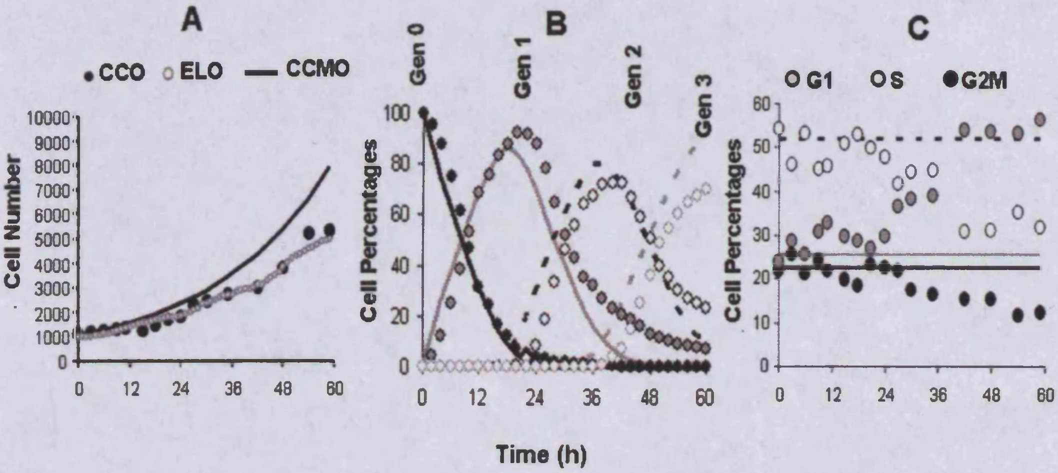
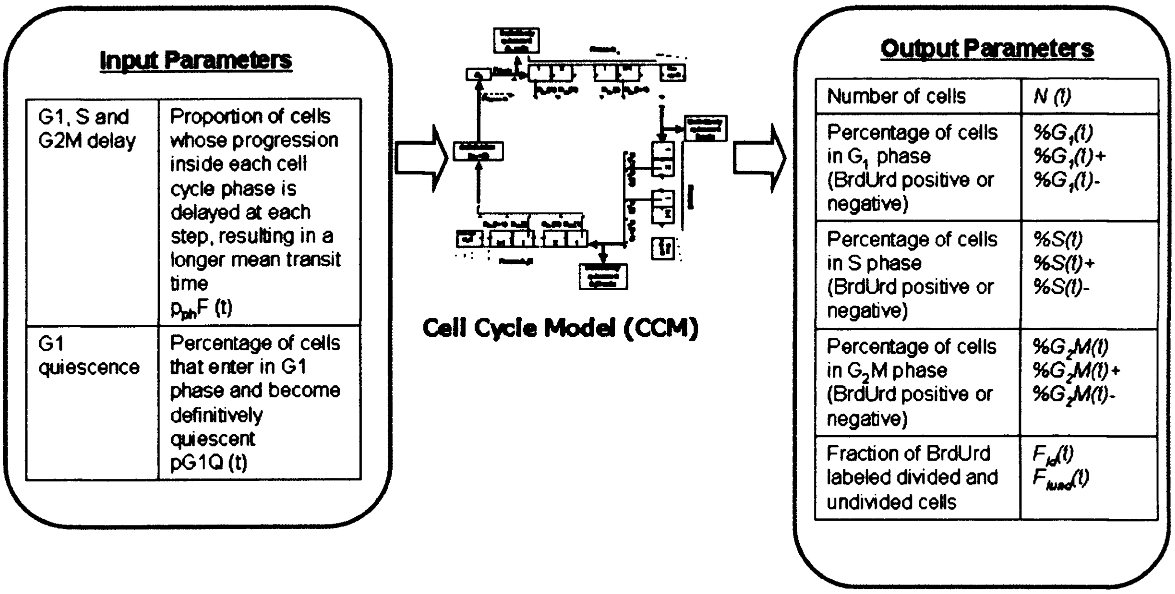


Fig. 6-6 ***Updated comparison.*** *With new CV values the model was simulated again and was compared with experimental data as described in context of figure 6-3.*

6.3.3 Variable phase durations and cell cycle perturbations - optimization of the cell cycle model

Considering all the results shown in the previous two sections, it became apparent that the outputs from the different cytometric platforms overlapped sufficiently but a further level of complexity needed to be considered for full integration. The parameter outputs from the encoded lineages (direct and calculated) provided an independent method for constraining and defining the cell cycle model parameters. The aim therefore was to achieve the best-fit against all experimental data and hence provide an independent multi-dimensional approach for cell cycle model optimization. The outputs from the CCM made no *a priori* assumptions and could be considered as the 'raw' outputs. To determine the extent of 'good fit' between the model and experimental data outputs, sum-of-squares error (SSE) were obtained (see Appendix XIV) and reported in Table 6.3.

As highlighted from FC data and from EL data, even control cells growing in normal conditions were characterized by physiological phenomena that contributed to partial synchronization of the cell population. These were particularly evident if we took into account that cell phase distribution over time, a completely asynchronous cell population should be a flat line, whereas in this case a slight oscillatory trend was evident for the experimental data (FC data) in figure 6.6 C. This oscillatory trend was also evident in figure 3-4 C where local slope of the normalized cumulative growth curve for control condition was measured. As described before these oscillation represented the wave of cells delivering to mitosis for each generation; and the sequential decay of the response reflected that population growth tends towards asynchrony. Cell manipulation and quiescence factors probably determined the partial accumulation of cells in G1 phase at long times of observation, represented two of the effects observed even in control samples. In particular, it was decided to represent the effects of cell manipulation as a delayed progression through G1, S or G2M phase for undivided cells or for cells that have divided only once. With the passage of time, cells progress through the cell cycle and the effects of the manipulation tend to disappear leaving place to the creation of a subpopulation of quiescent cells. Considering all these assumptions, a new set of input parameters were introduced and the cell cycle model was optimized against experimental ELO, FCO and CCO.



Input Parameters

G1, S and G2M delay	Proportion of cells whose progression inside each cell cycle phase is delayed at each step, resulting in a longer mean transit time $p_{ph}F$ (t)
G1 quiescence	Percentage of cells that enter in G1 phase and become definitively quiescent pG1Q (t)

Output Parameters

Number of cells	$N(t)$
Percentage of cells in G_1 phase (BrdUrd positive or negative)	$\%G_1(t)$ $\%G_1(t)+$ $\%G_1(t)-$
Percentage of cells in S phase (BrdUrd positive or negative)	$\%S(t)$ $\%S(t)+$ $\%S(t)-$
Percentage of cells in G_2M phase (BrdUrd positive or negative)	$\%G_2M(t)$ $\%G_2M(t)+$ $\%G_2M(t)-$
Fraction of BrdUrd labeled divided and undivided cells	$F_{ld}(t)$ $F_{lund}(t)$

Fig. 6-7 *Updated input parameter estimates to optimize the CCM describing the unperturbed growth of U-2 OS cells.*

The set of parameters that enabled a good simulation of the experimental data derived both from EL and FC were determined using a non-linear fitting procedure. The constrained non-linear fitting was accomplished with the 'Solver' function associated with the MS Excel spreadsheet. The Solver function was based on the Generalized Reduced Gradient (GRG2) algorithm. This procedure enabled the optimization of the following parameters: G1, S and G2M delay rate for generation 0 and 1 and G1 quiescence probability for generation 1, 2 and 3 or more. In the following table, the final results of different attempts have been reported.

Table 6-2 *The parameter matrix used to optimize the CCM*

	G_1 Del	**S Del**	**G_2M Del**	**G_1 Del +**	**S Del +**	**G_2M Del +**	**G_1 Q (+&-)**
Generation 0	**0.37**	0.00	**0.18**	0.00	0.00	0.00	0.00
Generation 1	0.00	**0.23**	0.00	0.00	**0.14**	0.00	**0.15**
Generation 2	0.00	0.00	0.00	0.00	0.00	0.00	**0.18**
Generation 3	0.00	0.00	0.00	0.00	0.00	0.00	**0.38**

G1 Del, S Del and G2M Del imply that a proportion of cells progressing inside each cell cycle phase were inhibited at each step, resulting in longer mean transit times. For instance: (i) G1 Del=0.37 meant that cells in generation 0 need a longer time to pass through G1. With this particular value of the delay the mean duration of G1 is about 5.4h. (ii) G2M Del=0.18 meant that cells in generation 0 presented a mean G2M duration of about 7.2h. (iii) S Del=0.23 or S Del+=0.14 meant that cells in generation 1 progress through S phase in 13.4h or 12.0h respectively. The symbol "+" meant that the parameters were related to BrdUrd-positive cells, when not specified we refer to data derived from BrdUrd-negative cells. G1 Q (+&-) represented the percentage of cells that, once entered in G1, became definitively quiescent.

To fit the cell number and flow cytometric percentages the Solver function was imposed to minimize the SSE, for cell cycle percentages, for absolute cell number and for cell generation distributions.

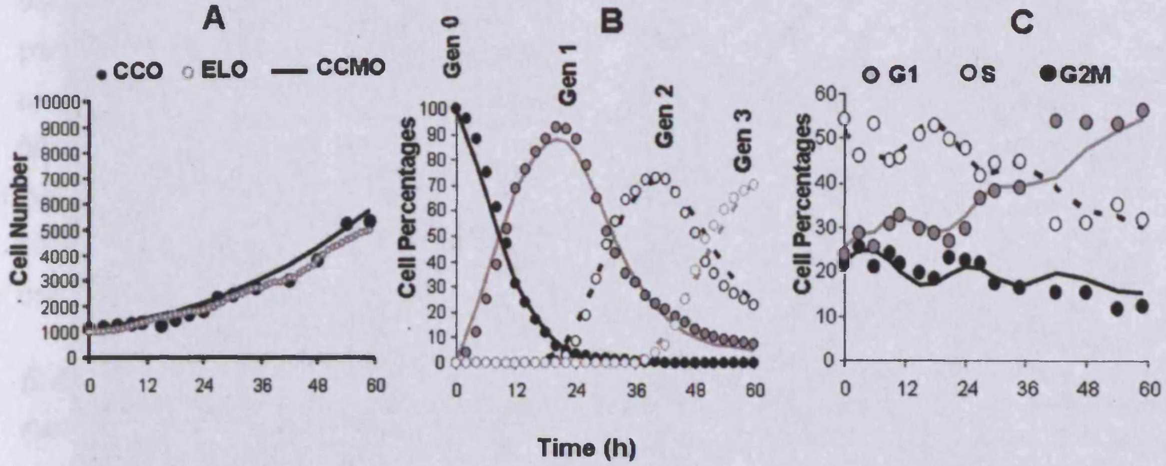


Fig. 6-8 *Comparison between simulated (continuous line) and experimental data (symbols). The simulation was obtained setting the parameters to the values listed in table 6.2.*

A good correlation between model output and different cytometric platforms were obtained for all considered data. The SSE values after optimization compared to previous SSE values provided an overview on the data fit improvement (Table0.3).

Table 6-3 *SSE comparison before and after optimization.*

SSE (Only considering inter cell variability)	**SSE (After optimization)**	**Comment**
1.439 x 10^7	1.284 x 10^6	Between Coulter Counter and cell cycle model outputs
6.429 x 10^3	6.047 x 10^2	Summed SSE for each cell cycle phase between FC and cell cycle model outputs
5.075 x 10^7	1.710 x 10^6	Summed SSE for each generation between Lineage and cell cycle model outputs

Cell cycle phase duration with respective coefficient of variance remains same as the previous simulation as presented in Table 6.1. These values, with the additional parameters describing the deviations from asynchronous growth (delays and quiescence parameters) enabled a complete fit of the data against all experimental platforms, and as such characterized the cell cycle of the control population in a platform independent fashion. This ability provided the advantage to utilize these parameters as platform independent units to measure cell cycle dynamics, and to exemplify such advantage, cell phase boundary (CPB) values obtained through the model were exploited to measure cell age distribution within the lineage data encoded by ProgeniTRAK.

6.4 Using the cell cycle phase boundary parameters to measure cell age

The aim of the following section was to assign cell cycle age to cells encoded by ProgeniTRAK by using the cell phase boundary (CPB) values or cell cycle phase duration parameters obtained from the cell cycle model. In a heterogeneous cell population at any one time the cells within the population consisted of cells at different phases of the cell cycle.

If we consider a cell featured within a lineage map, we know a limited amount of cell cycle information such as the time from the last mitosis and time to the next mitosis. The effective age therefore of the cell at any particular time point was defined as the time elapsed since last mitosis and therefore designated as absolute cell age (age_{abs}). Cell age can also be defined as a relative term to the time elapsed since the last phase boundary was crossed. In other words in the case of determining relative age, we reset as a cell crosses each phase boundary, therefore we can identify two transitions: G1 > S-phase, S-phase > G2M phase, providing parameters designated as age_{G1}, age_{s}, age_{G2M} respectively. It is important to note here that even though M was considered a separate phase from G2, due to the short duration M was included with G2 and designated as G2M phase, also flow cytometry approaches cannot distinguish between G2 and M.

To exemplify absolute age and relative age, let us consider a $cell_X$, if $cell_X$ first appeared (from the mitosis of the progenitor cell) at 32.7 h after the start of the experiment and divided into two daughter cells at 52.8 h, the $cell_X$ would have a complete IMT value of 20.1 h. Therefore when $cell_X$ was observed at 45 hours after the start of the experiment $cell_X$ would have an age_{abs} of (45-32.7) = 12.3 h and age_s of 8.4 h; age_s was designated because according to table 6.1 the cell should have crossed the G1 > S boundary 3.9 h after mitosis at 32.7 h. So at (32.7+3.9) = 36.6 h the cell entered the S phase of the cell cycle and also at that time the relative age was reset to 0 h and accordingly at 45 h, the cell had an age_s of (45-36.6) = 8.4 h. Again if $cell_X$ were viewed at 50 h, the reassignment would be age_{abs} = 17.3 h and age_{G2M} = 3.1 h respectively.

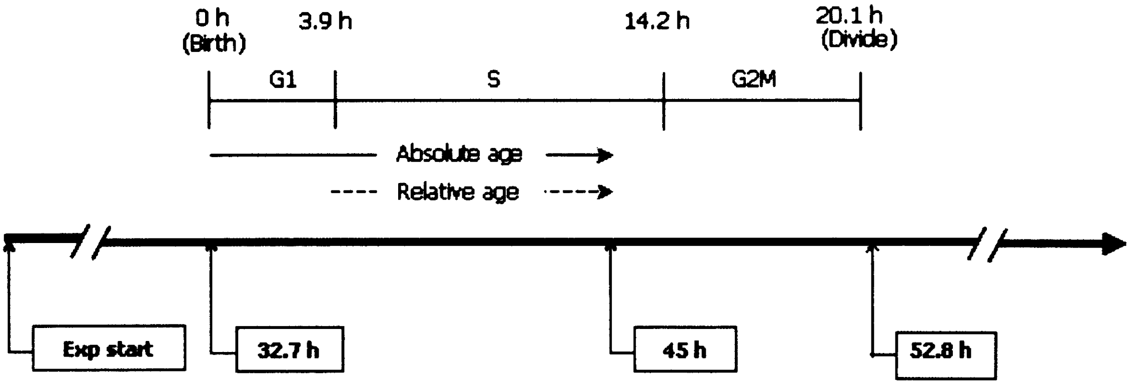


Fig. 6-9 *Diagram to illustrate the designation of absolute and relative cell age.*

So the relative age of $cell_X$ consisted of three attributes: (i) the age (in h) (ii) the cell cycle phase (iii) and generation information. As stated earlier since the EL and CCM only included the generation information, data from these two sources could be utilized for investigating the cell age. Another aspect for measuring cell age was the variation of cell cycle duration. In normal culture conditions not all cells have the same cell cycle duration IMT or Tc value of 20.1 h (see figure 6-5) inter cell cycle time variability was the prime cause for this phenomenon and the variation was innate and remained the primary source for population asynchrony and heterogeneity. Different aspects of this variation of cell cycle duration (IMT) in control conditions were analyzed in chapter 4. From the CCM the variation was artificially imparted by covariance indices and thus the output population also incorporated the heterogeneity of total cell cycle time, the mean CPB values as outlined in table 6.1 were G1 > S 3.9 h and S > G2M (3.9 + 10.3) = 14.2 h respectively. The objective of this section of the analysis was to exploit these CPB values to dissect the IMTs encoded through ProgeniTRAK and FluorTRAK and thus explore the implications of using model derived parameters to transform experimental data.

During the first stage the IMTs encoded through ProgeniTRAK were subjected for analysis, for each encoded cell the standard CPB values generated earlier were readjusted to define the cell specific CPB by a simple method. For example if the IMT for a particular cell was found to be 30 h, the CPB values would be readjusted to 3.92 x (30/20.1) = 5.9 h and S>G2M boundary to 14.2 x (30/20.1) = 21.2 h respectively. Through this simple readjustment the delay as well as early IMT was attributed evenly to all phases of the cell cycle, which arguably might not be biologically accurate as deviation from average cell cycle time (20.1 h) can be the product of many factors as implied in table 6.2 and would most likely differentially effect G1, S-phase and G2 duration. However, despite this limitation each encoded cell that delivered to mitosis were subjected to cell age measurement reviewed at 6 time windows with a 9 h interval from time 0 to 54 h (e.g. at 9, 18, 27, 36, 45 and 54th h). At each time interval the first step was to index cells that could be included into the analysis, the criteria for a cell to be indexed was that the cell must have its first appearance before the time point under consideration as well as delivering to mitosis after that time point. A Perl script (analytical tool) was written that would determine these criteria automatically for each encoded cell. Once indexed the next step for the analytical tool was to measure the three attributes of relative age taking into account the readjustment of the CPB values. According to these attributes cells were binned to the nearest hour bin and plotted as a distribution. Each plot represented the distribution of cell age for a particular time point and generation.

For the CCM the output automatically index each cell at each time point with its generation, phase and relative age information and as such were binned and plotted with the same graph as that of EL data. For each graph the smooth line represented the CCM output while the noisy line represented the EL data (CCM had substantially higher number compared to EL).

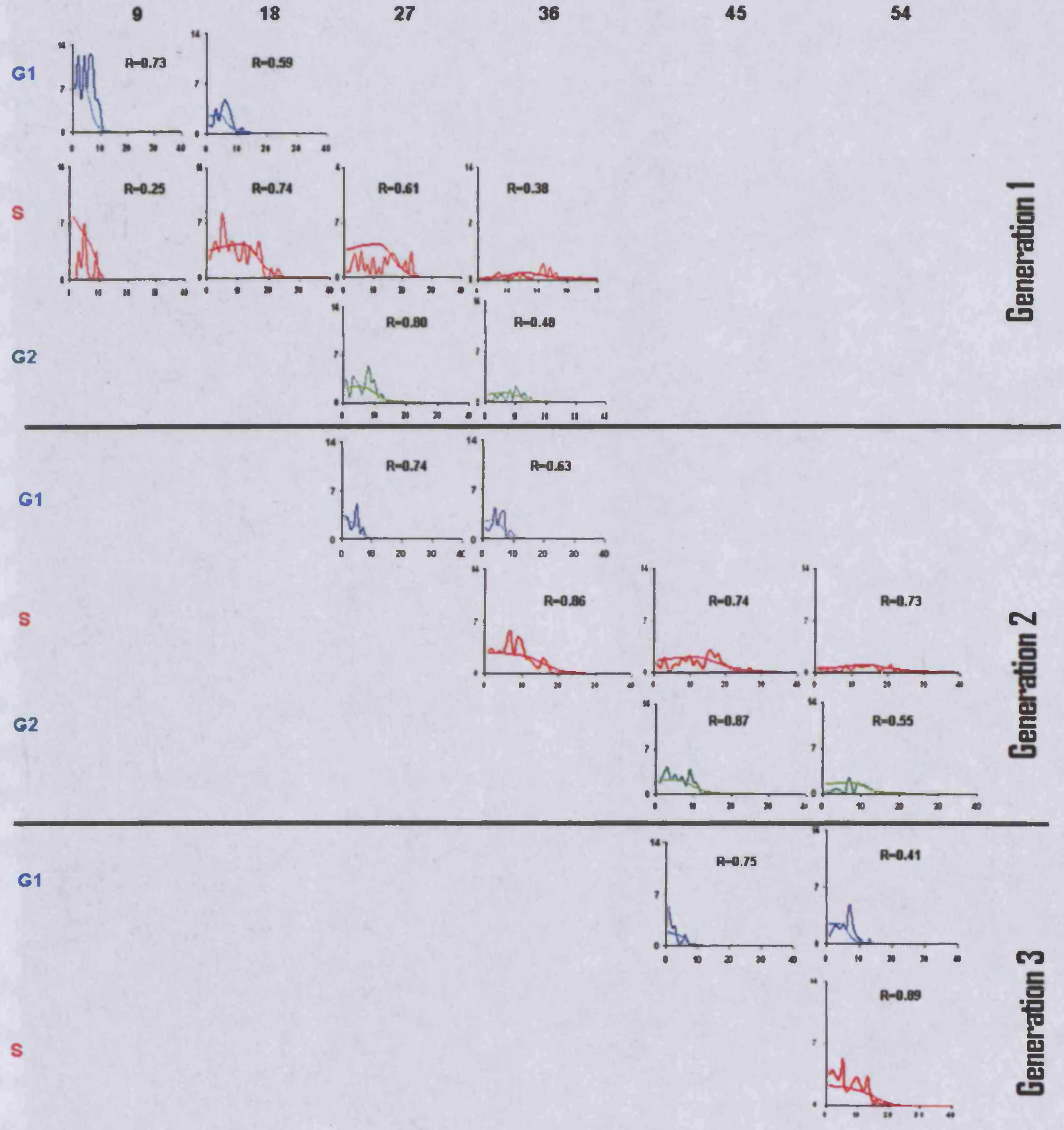


Fig. 6-10 ***Cell age distribution.*** *Progressive graphs showing cell age distribution at specified time points (9,18,....54 h) for generation 1, 2 and 3. For each generation, the relative age distribution were segmented according to the phase of the cell cycle (G1 blue, S pink, G2 green). With each graph the smooth line represents cell age distribution calculated from CCM while non-smooth line represents EL data. The Pearson's correlation value between the two distributions was inserted into each graph.*

Figure 6-11 is a typical representation of the population dynamics where a population of cells undergoes cell cycle traverse and accordingly moving from one phase of the cell cycle to the next. For example, at 9 h after the start of the experiment, all the cells of generation 1 were assigned to G1 or S phase of the cell cycle (predominantly in G1 phase) but no cells were in G2M. It is important to note here that progenitor cells (generation 0) could not be included in the analysis since the respective IMT was not available (reason explained in chapter 4). At 18 h much of these cells that were in G1 at 9 h moved to S phase but cells that belong to S phase at 9 h had not yet entered the G2M phase. At 27 h, all cells that were in G1 phase at 18 h moved to S phase while cells that were in S phase at 18 h also moved to the final G2M phase. Also at 27 h some of the cells completed the first cell cycle and yielded the next (2nd in this case) generation of cells and indeed were in G1 phase. However as a small number of cells entered to the 2nd generation at 27 h the age$_{G1}$ distribution was small as observed in the graph. At 36 h the cell population was very much heterogeneous as they encompassed not only cells from generation 1 and 2 but in terms of phase they were distributed across all three phases of the cell cycle. However if the matrix of graph in figure 6-11 was viewed in columnar fashion, i.e. at certain time points, it would be possible to distinguish a certain sub-population of cells that were predominant for that time point. For example at 36 h a heterogeneous population of cell belonging to both generation 1 and 2 and all three phases of cell cycle were observed, but the predominant sub-population of cells belonged to S phase of generation 2.

Acknowledging the cell number difference between CCM and EL data, the distribution of relative cell age was compared with a simple Pearson's correlation coefficient that measures the strength of the linear relationship between two variables. The positive Pearson's correlation value of each graph in figure 6-11 when coloured with 0-1 gradient scales, yielded a heat map where the maximum correlations (1.0) were shown with blue colour while minimum (0.0) with green colour.

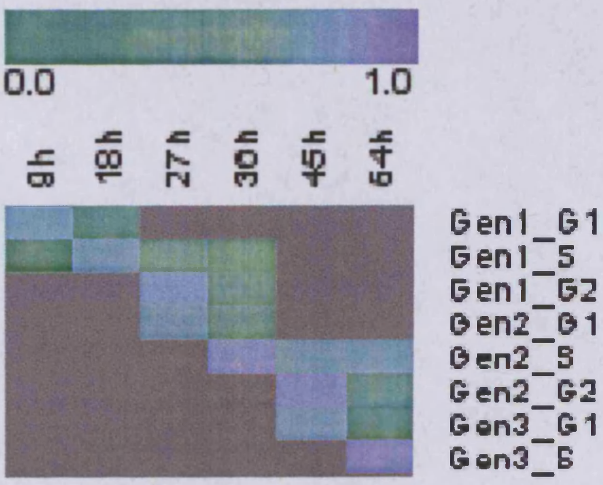


Fig. 6-11 *Heat map view of relative cell age correlation between model and EL data.*

This correlation heat map in figure 6-12 was a succinct visualization of the cytometric data integration process at the meta data level. A strong correlation was observed diagonally through time implying a synchronized population at this detailed resolution and also corresponded to previous results in figure 3-4 C, indicating a sub-population of cells cycling in a synchronized manner. The measurement and visualization approach assisted to assign generation and cell cycle phase tags to each cell in the lineage assays at a particular time point. For example if measured at 30 h via this schema, it can be stated that most of the living cells would belong to 1st generation and at the S-phase of their cell cycle and the distribution can be exploited as a signature of population heterogeneity and thus can be utilized to measure the effect of perturbation in population context. Moreover the matrix of progressive distributions demonstrated the population dynamics as it showed the passage of cells from one phase to another, as well as from one generation to the next, this time dependent population response would assist for better understanding the population dynamics in different experimental conditions.

6.5 Concluding remarks

We have investigated the idea of integrated cytometry and mathematical modelling. Cell cycle models at present aim to simulate cellular dynamics either through information about genes and protein networks involved in the cell cycle process (Alfieri et al. 2007), or through exploiting multi-parametric FC data that depict cell cycle kinetics and control mechanisms of the cell cycle (Pierrez and Ronot 2004), all cases the common goal has been to provide predictive models and consequently highlight the interactions and emergent properties of dynamic cellular systems. However, there is a consensus in the community (mathematicians and biologists alike) that cell cycle mathematical modelling has not yet reached this position (Sible and Tyson 2007), the principal reason being that there needs to be a better interweaving and convergence of experimental data, model derivation, simulation and validation. In addition, mathematical models are predominantly validated using the same data source as that used to develop the model (a cross-validation approach) (Bertuzzi et al. 1988; Sible and Tyson 2007). In the current work we suggest that the most plausible approach to overcome such limitations is through integrating data from different sources (i.e. acquisition platforms). We have shown that the cross-platform approach offers bilateral benefits: the compartmental cell cycle model (derived from flow cytometry data) obtained new constraints imposed by microscopy and Coulter Counter data (see Appendix XV). Equally, the optimized parameter outputs from the cell cycle model assisted to dissect and enrich the lineage maps – this convergence provided new insights to progeny responses where the delivery to mitosis by a synchronized subpopulation of cells were identified. The next stage is to meet the challenge of building a continuous (not statistical) mathematical model that can contend with cell division nodes (bifurcation) within a lineage map, incorporating aspects of inheritance and memory. We are in a position to integrate cytometric outputs (cellular and molecular) to meet this objective. Essentially, the comprehensive data rich information addressing dynamic cellular systems could be considered to be scattered across many cytometric platforms; establishing a robust and far-reaching data integration process would meet the hitherto unmet need of multi-scalar large data acquisition to underpin cell cycle mathematical models (Systems Biology Report 2007).

Chapter 7: Discussion – impact of cellular informatics

The study presented in this thesis has established a prototype informatics infrastructure that enabled encoding, archiving, mining and the interpretation of timelapse image data comprising cellular behaviour. The cell lineage map provided the principal component of such an infrastructure representing the pedigree structure within the cell population. Here the cell was considered as an object and the object behaviour was influenced by many different factors including genetic and epigenetic influences on programmed behaviour; as well as external influences including environmental stress and others. The lineage map introduced through this research aimed to encapsulate the complexity of the proliferative responses and variability within populations associated with divisional events, and at the same time provided a generic framework for the application of bioinformatics techniques for both data management and information retrieval.

Previous studies have been carried out to depict and interpret cellular lineage behaviour predominantly with respect to stem cell dynamics (Braun et al. 2003; Dzierzak and Speck 2008; Geard and Wiles 2005; Karam 1999; Orkin and Zon 2008). Few efforts were also made in other somatic cell context (Chu et al. 2004; Endlich et al. 2000; Forrester et al. 2000; Forrester et al. 1999; Prieur-Carrillo et al. 2003), but these efforts have not encompassed the inheritance-based relationship within a population context as well as inter lineage or inter-nodal relationships, and therefore was not sufficient to depict intricate spatial and temporal dynamics of cells. Additionally, these encoded cellular behaviour parameters were not systematically linked with experimental descriptors and thus undermine the hypothesis driven data mining at the metadata level. This research attempted to overcome some of the limitations by introducing a novel data format which underpins the investigation and established not only intra and inter lineage relationships, but relationships substantiated with experimental descriptors. Such intricate relationships enabled metadata level data mining and subsequent facilitation of a better understanding on innate and acquired cellular processes and also revealed occult relationships that are separated over a wide time window, for example the influence of a topoisomerase inhibitor on progenitor cells and the overall impact on later generations. Another important aspect of the lineage data format was that it supported the visual representation of the progeny tree during and after the encoding process. Visual representation of the cell lineages, contrary to actual cells in the image sequences, can be correlated with visualizing the annotated genome map instead of the single DNA. A gene sequence even with detailed annotation gives a poor understanding from a genome perspective, but when represented within a genome map (see: NCBI Map viewer http://www.ncbi.nlm.nih.gov/projects/mapview/maps.cgi?taxid=9606&chr=1), the same annotations produce context sequence information. Thus the data format and the lineage structure simultaneously provided a computer readable map to explore the cellular behaviour through hypothesis driven data mining and a visual map for human interpretation as illustrated by the figure bellow.

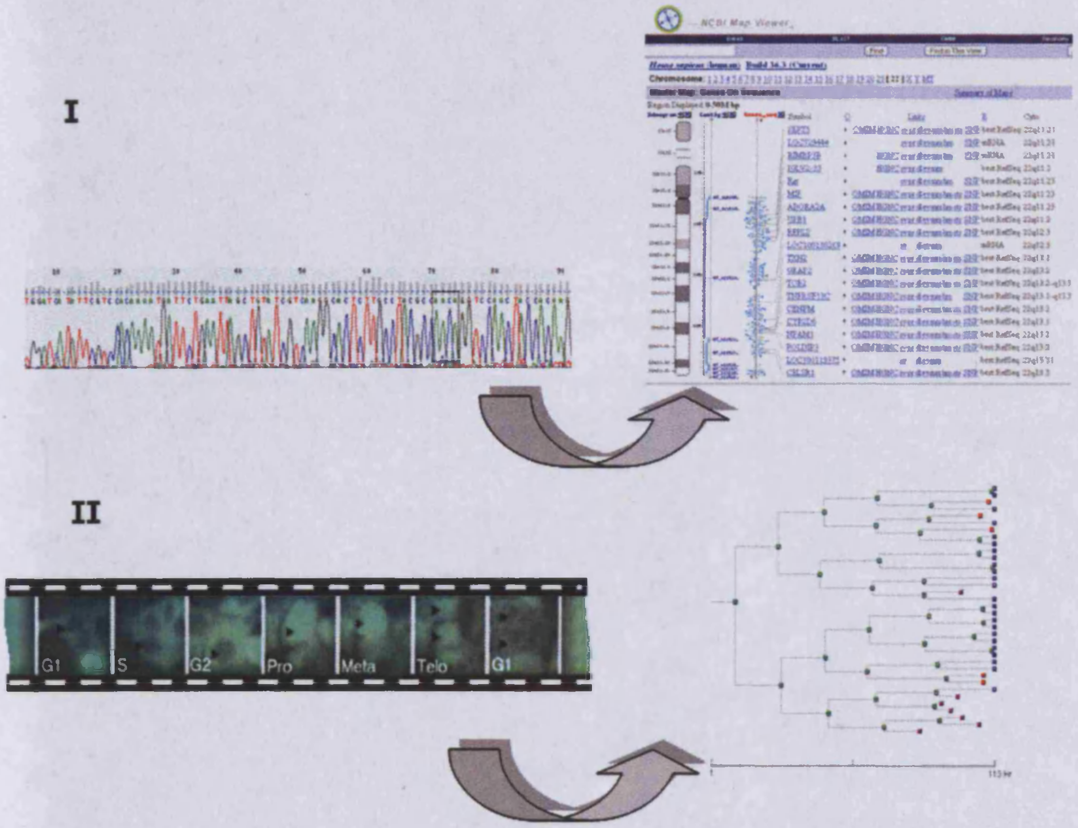


Fig. 7-1 ***Visual representation of encoded data.*** *Panel I shows how a lower order DNA sequence were organized and given a visual representation in the NCBI Map Viewer within a genome context. Panel II shows how a cellular behaviour image sequence has been given a visual representation through lineage map encoded via ProgeniTRAK/FluorTRAK.*

Current software packages of both commercial and academic origin have improved our ability to convert "images to parameters" – a goal shared by many disciplines including cell biology but subsequently has introduced new challenges for managing and mining these image derived numbers and directly translating the cellular behaviour to *in silico* models. Through a prototype infrastructure, the current research established encoding, archiving, mining and visualizing tools. The encoding process or tools encapsulated the critical features of cell-cell heterogeneity, molecular dynamics, phenotypic behaviour and time-dependent events. The multi-level descriptors and parameters attributed to each cell, or node within the resultant cell lineage maps, provided a unique framework for applying bioinformatics-like query algorithms. Further, the ability to locate molecular profiles of cell cycle phase traverse and checkpoint responses was also achieved by encoding high-resolution fluorescence image sequences (GFP-cyclin B1) in a lineage format. The two levels (experimental and single cell level) nomenclature embraced the ontology hierarchy as described in Fig 7-1 and a visual representation of the parameters at the experimental level facilitated users to select experimental parameters without manually typing the cell name. This process not only oriented the user within the experimental scenario but also provided an error free tagging of the lineages. At the single cell level the evolving lineage map with its associated colour tagged nodes oriented users to encode the lineages in an accurate manner and at the same time associated the image derived parameters to the correct single cell, node and hence lineage branch. This was important since visual orientation of complex yet dynamic data reduced the possibilities of error of encoding and provided the foundation for future automation. The encoding process embraced the document writing process such as that presented in everyday packages such as MS Word, where the visual representation of an evolving document was provided and at the same time has the features to editing, saving of the present and previous document. Both in ProgeniTRAK and FluorTRAK this philosophy of encoding was adapted, which indeed made the encoding process flexible and "user friendly".

Archiving the encoded data became the next important feature and provided the fundamental shape of a developing infrastructure, therefore initially the decision was made to undertake a primary archive in tab delimited text files. There were three reasons for adopting a text file format – first, the lineages were encoded following the philosophy of MS Word, where each document was saved as one file in the folder and likewise each lineage was saved as single text file in the "Lineage folder". It is important to acknowledge here that within this general philosophy, lineages could be encoded in an XML format but defining the XML tag and constructing the associated XML parser, was not essential at this stage of development since the main focus of the research was to demonstrate the future prospect of encoding cell lineages from timelapse image sequences and exploring the relevancy and usefulness of these lineage derived data for hypothesis driven data mining. The second reason for selecting a text file format was due to its simple editing facilities (e.g. removing or introducing new column of data), which was extensively utilized during the developmental process. Third and finally the text delimited text file gave an extra advantage in terms of data mining and analysis as it was the simplest format that was easily readable by all common computer languages. Hypothesis driven data mining is always an ongoing process and therefore demanded new solutions at different time points. A number of PERL scripts were written according to the hypothesis or idea and since the data source was in text file format, the process of data mining was simplified.

One major limitation of text file format was data sharing, even though the "Lineage folder" which contained all the linage files were situated in the shared drive, but only limited users had access to the data i.e. locally-based. In order to provide multi-site accessibility, a web accessible database was introduced and chapter 3 illustrated the process of implementation and mining such database. ProgeniDB (Khan *et al.* 2007) provided event based data and from the perspective of complexity, this database may be categorised as a simple database, however this prototype database demonstrated multiple benefits of having public access encoded cellular behaviour database contrary to an image database (Marx 2002). The primary benefit was data reduction as image files consume more memory contrary to alphanumeric encoded data, so the ease of archiving and sharing data became very efficient (data reduction ~100 GB to ~100 KB). However the true benefit lies with the fact that encoded data did not require subjective interpretation, which makes encoded data readily interpretable and shared to cell biologists, mathematicians and statisticians alike.

In the near future relationship derived data like inter mitotic time and distance (i.e. IMT, IMD) will also be retrieved from ProgeniDB as new data mining interfaces will be introduced. At present ProgeniDB marked the completion of the infrastructure, even though at a small scale; the infrastructure demonstrated the possibilities and usefulness of cellular informatics. A major limitation of the encoding process was indeed the absence of cell tracking automation – at present the encoding process can be categorized as a semi-automated process where cell tracking and event recognition were performed manually. Tracking cells after the bifurcation point within an image sequence, remained the overwhelming challenge and past decades have witnessed several attempts to resolve this challenge (Bao et al. 2006). Even though set out as a clear requirement and specification, the development of automated image processing algorithms were beyond the scope of this thesis. Tracking the cell object acquired in phase contrast transmission was not trivial, but it was clear that for the cell lineage approach to work in a high-through-put mode a better cell tracking approach has to be addressed. Therefore, the need for automation and a large dataset was acknowledged as the primary bottleneck of such an endeavour. Currently described image analysis tools do not contend adequately with the mitotic event particularly upon the generation of two daughter cells (bifurcation) (Braun et al. 2003). The simple premise is that that more information maybe available by analysing the mitotic event within a timelapse sequence using forward and backward filters (Hamahashi et al. 2005), the collaborative work aims to constitute an enabling technology providing capacity in data generation.

The metadata provided the means for a detailed analysis of encoded image data. Detailed analysis and interpretation of the data outputs were presented in previous chapters. Here the discussion covered the breadth and depth of such analysis; the primary reason was to investigate the benefits of extracting such dynamic outputs in fulfilling the knowledge gap that exists between molecular level and systems level information as shown in fig 1-2. Three cellular levels of complexity as outlined in the fig 1-2 were addressed through the lineage data format.

7.1 Single cell level analysis

Analysis at the single cell level involved segmenting the lineage map in a vertical fashion, perpendicular to time, generation or both. The outputs of such a segmentation comprised population distribution behaviour that encompassed events such as IMT, IMD. These distributions when grouped according to time, generation or both generated a new level of knowledge particularly with respect to IMT as this parameter could not be directly quantified from the image sequences and as such has not been included in previous conventional analytical procedures. Cell cycle related parameters like IMT were designated as ‘on-the-fly’ derived parameters and exploited the inter-nodal relationships and patterns established through this lineage data format.

Even though these parameters and encoded events were regarded as discreet data points, but linked with multi-dimensions such as time, generation, drug, dose a comprehensive analysis was performed that ranged from common event curve to complex drug and dose dependent GFP-cyclin B1 profiling. These profiles not only generated detailed knowledge as they underpin the dynamic molecular process, but also provided the foundation for developing mathematical models as they encapsulated up to three tractable molecular profiles simultaneously. Even though the encoding process is time consuming and requires a high level of expertise for indexing the 3 ROIs at each time point (nucleus and cytoplasm in case of experiment involving GFP-cyclin B1). Due to this limitation a small number of lineages were encoded using FluorTRAK, but the implication of even a small number of lineages was emphasized by the publication of two peer-reviewed papers (Chappell et al. 2008; P´erez-Vel´azquez et al. 2008).

Encoded lineages also provided spatial data (cellular movement or motility etc) that can be retrieved from the lineages. The impact of these data were explained in terms of colony formation, directionality and motility of the cell. Further implications of these attributes were not investigated to the full potential as these data are more pertinent to wound healing research. However a collaborative pump-priming project has been initiated with Dr. Patricia Martin at Caledonian University, Glasgow to develop a modified version of encoding tool – WoundTRAK to encode cellular behaviour in the context of a primary cultured 3D wound model.

7.2 Lineage level analysis

Both ProgeniTRAK and FluorTRAK established the relationship of cellular dynamics at the lineage level. The unique nomenclature of cells with ‘B’, ‘BS’, ‘BSN’ etc. provided a tagged nomenclature to maintain links between all progeny. IMT and IMD were the two most important inter- and intra- nodal readouts that were highlighted through the data format while maintaining relationships. IMT or cell cycle time remained the fundamental readout for describing the pharmacodynamic response that was a prime pre-requisite for the process of drug screening. The variation of IMT and its relationship with drug (derivative, dose, bolus) was measured from different perspectives – in terms of generation-to-generation (vertical segmentation) and in a sequential manner (horizontal segmentation). The lineage structure when analyzed vertically provided the opportunity to quantify the distribution of IMT as per generation and in relation to drug parameters. When lineages were segmented and analyzed in a horizontal fashion, the cell cycle action of drugs could be quantified in a pedigree structure, which provided the means to quantify how the effect of the drug was cascaded from one generation, along with the asymmetric distribution of the effect between two daughter cells. These results generated from exploiting the relationship basis of the lineage or pedigree format, form the basis for developing predictive models. These predictive models could then be utilized to identify the cells that in future would generate the resistant progeny and if such identification were possible with certain confidence a rare opportunity to explore the genetic profile would be possible. Such a hypothetical situation would certainly contribute significantly to our understanding about the origin of resistance at the molecular level.

Through a lineage analysis approach a simple yet potential bioinformatics-scoring schema was introduced to score the track within a lineage. This theme for such a scoring schema evolved from the classical bioinformatics algorithms like BLAST and aimed to establish a comparative cellular behaviour algorithm, where cellular behaviour from a particular experimental condition (gene knockout, drug treated etc.) could be compared and scored against normal cellular behaviour, similar to that of an unknown sequence aligned against well annotated sequences to gain preliminary knowledge about the unknown sequence. This bioinformatics approach will benefit the process of drug discovery as PD response of a new chemical agent (NCE) can be scored and aligned against already known PD responses.

7.3 Multi-cellular level analysis

At this level of analysis heterogeneity of cellular behaviour within a population was analyzed. The hypothesis was that that the asymmetry of cell division time (IMT) provided the primary route for heterogeneity and comparative analysis of IMT between daughter cells figure. 4-10 reflected the consequences of drug treatment on the symmetry of the bifurcation. Since the relationship of cellular behaviour embedded in a lineage map can be explored from different perspectives as outlined in chapter 4, it was possible to integrate this multi-cellular heterogeneity not only to understand the tumour resistance but also wound healing, cell migration, cell attachment, cell proliferation, angiogenesis, senescence and other cellular processes. However, to achieve such ambitious objectives the primary criteria was to demonstrate the pertinence of such a bioinformatics infrastructure.

7.4 Assening the generic applicability – indicators for future sustainability

Present applicability and future growth of any research depends on the generic nature and flexibility for improvement, probably this is even more important for bioinformatics research (bioinformatics is research not a 'cottage industry'). Our collective bioinformatics endeavours encompassed both research and development; where research included delivery of novel algorithms, data format while development included implementation of tools and database. The overall activity addressed a wide range of yet unmet intellectual challenges that enveloped a spectrum of research interests and associated demands for tool development. Bearing in mind the multi-faceted nature of the current study; the thesis undertaking demonstrated its generic nature and future development potential, while addressing the hypothesis and objectives set forward. The hypothesis that spatiotemporal cell kinetics data in a lineage format provided an essential route to determining cellular dynamics is not only relevant to cancer research but showed to have shared benefits in all types of cell based research, like wound healing and senescence studies. A growing interest of using such an informatics infrastructure was convened during this research period and ProgeniTRAK was used in context of wound healing with Dr. Patricia Martin and demonstrated a considerable knowledge uplift, which has led not only to publications (manuscript under preparation) but also grant application (submitted with Dr. Patricia Martin, Glasgow Caledonian University) to formalize the future opportunities.

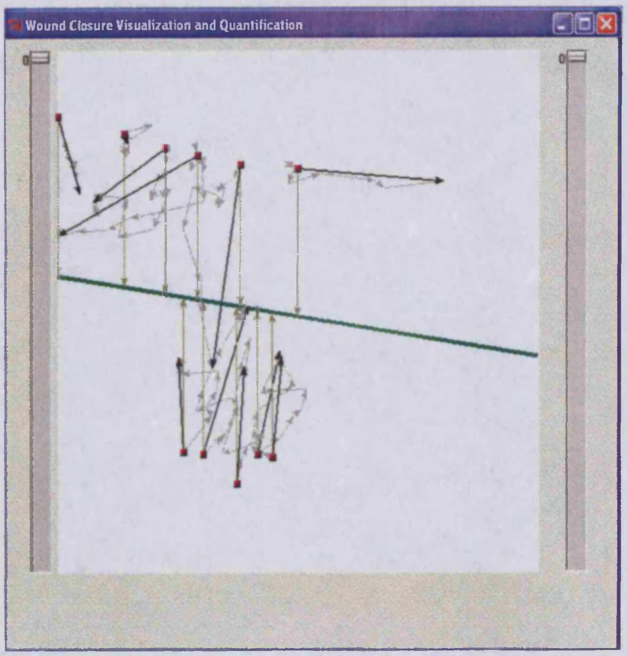


Fig. 7-2 ***A prototype GUI showing the cellular dynamics of wound healing****. Gray lines show the actual movement of the cells from its origin. Black lines showing start to end point distance travelled and Cyan lines showing the minimum distance the cell should travel towards the wound bed.*

Recently the infrastructure was also explored in several other contexts such as search for new drugs for late G2 checkpoint control (with Prof D Kipling, Pathology, Cardiff University); and more recently has been applied to work carried out with primary mouse embryonic fibroblasts (MEF) (Prof. Alan Clarke, Biosciences, Cardiff University).

Aside from these rolled out collaborative efforts, the most important testimony of the generic nature of the bioinformatics framework stemmed from the cytometric data convergence as discussed in chapter 6. Such convergence of cross sectional flow cytometric data with timelapse data not only reinforced the generic nature of this data format and the infrastructure as a whole but also showed the potentiality of formulating a comprehensive knowledge base for cytome behaviour that will be crucial for metadata underpinning of systems biology and drug discovery.

7.5 Cellular informatics – systems approaches to biological research

The aspiration of bioinformatics is to provide the operational and intellectual framework for a systems understanding of biology towards predictive *in silico* models. Although systems biology is in its relative infancy (Kitano 2002), the potential benefits are enormous in both scientific and practical terms. Systems biology invokes an interdisciplinary research community to develop detailed mathematical models to simulate cell regulation, pathways and interaction networks for molecules to provide systems level insights (Gibbs 2000; Noble 2002b; Sander 2000). Such models may help to identify feedback mechanisms that offset the effects of drugs and predict systemic side effects (Kitano 2002). Addressing such need, modelling "environments" are spawning that contain suites of tools necessary for model building, simulations, data fitting, and data management (Sible and Tyson 2007). Examples of such environments includes: Gepasi (Mendes 1993; Mendes and Kell 1997), Virtual cell (Moraru et al. 2002), JigCell (Allen et al. 2001) etc. In each of these environments there are tools to translate models into the Systems Biology Markup Language (SBML), a grammar that is becoming widely adopted by the biochemical network modelling community to exchange models (Sauro et al. 2003).

These endeavours indicate the ambitious transition occurring in biology from the molecular level to the systems level and as outlined in the introductory chapter, the complexity of dynamic cellular behaviour has remained at the centre of this transition process. At one end of the overall scale the purpose is to connect the nature and probability of a cellular response with the molecular networks that control such responses. At the other end, where the construction of mathematical models link multi-scalar analysis, enabling the undertaking of rational predictions. Models serve many purposes leading to the understanding of systems behaviour and prediction of complex responses to perturbation. However prediction poses a significant challenge since the molecular profiling of cell populations rapidly looses resolving power and value once a perturbing influence affects population dynamics (Lahdesmaki et al. 2005). The solution is to understand through modelling how molecular interactions influence cellular dynamics to reveal the confounding aspects of temporal responses and heterogeneity. Conversely, the disassembly and drill-down achieved through modelling can be used to identify informative cells so that molecular and functional studies can be applied to defined arms of a complex response or targets for new drugs. Once a model has been developed to an appropriate level of complexity, it can be run repeatedly and function as a high-throughput hypothesis platform (Endy and Brent 2001; Tomita 2001).

Considering the mammalian cell cycle engine as the primary descriptor for cellular dynamics; multi-level cell cycle modelling is required and for a mathematical model of the cell cycle to hold credibility with the biology community, it should be of sufficient complexity to incorporate a minimum number of processes known to be involved in cell-cycle regulation, including growth and division, growth restriction, survival, programmed cell death, DNA checkpoint control and cellular damage response mechanisms (Clyde et al. 2006). Although a great deal of knowledge of the biochemistry and the physical processes of the proteins that regulate the cell cycle was fairly recent (P´erez-Vel´azquez et al. 2008), mathematical models of the cell cycle can be traced back to as early as the 1970s (Hastings et al. 1977; Tyson 1974/75; Tyson and Sachsenmaier 1978). While considerable progress has been made in modelling methodology, and the discoveries emanating from the field of experimental biology are nothing short of remarkable, it is nevertheless the fact that, as things stand at present, there are very few quantitative models available, and none yet reached a level of accuracy and completeness required to engage effectively with translational research relevant to diseases of the cell cycle.

Multiple reasons contribute to this status of which the fundamental reason has been the poor understanding even from a qualitative perspective over the molecular drivers and their role to the process in a dynamic context. Quantitative and well annotated experimental data are required that identifies the molecular drivers responsible for the process, their heterogeneous distribution within a cell population and complex interaction in a dynamic context (Nelson et al. 2002b; Nelson et al. 2002c; Stirland et al. 2003). The development of methods for acquiring this quantitative knowledge is one of the greatest challenges for biology in the twenty-first century (Brent 2000). Even with the fulfilment of these ambitious requirements still the knowledge about cell signalling, feedback mechanisms, environmental effects and other parameters that define the cytome behaviour will be essential to formulate representative models for cell cycle.

What is currently lacking, however, is a unified approach to the problem and an organizational overlay that could ensure that relevant research is directed efficiently into a structured format suitable for multi-scalar modelling (Clyde *et al.* 2006). The bioinformatics infrastructure introduced through this research with its data format, encoding tools, databases, mining and visualization tools all provide a step change towards achieving this goal. The infrastructure was able to provide a highly sought aspiration of a multi-dimensional data format that can incorporate higher temporal information as well as cross-sectional data and furthermore allowed for the simultaneous quantitative comparison of protein expression and translocation. Recent publications of mathematical models using data generated from this infrastructure has been a testimony of such a longterm goal, where the first model (P´erez-Vel´azquez et al. 2008) represented an extended version of the transition state cell cycle model by Tyson and Novak (Tyson and Novak 2001) and has been linked with a model accounting for the inhibition effect of p21 on cyclin B1 (Pomerening et al. 2005; Pomerening et al. 2003), the second model (Chappell et al. 2008) included a novel coupling of both the PK-PD responses in context of an anti-cancer agent. This has provided new interlinked predictive models and goes along way to provide system level insights into the interactive mechanisms of drug targeting with the cell cycle (Systems Biology Report 2007).

As the pharmaceutical industry battles the escalating costs and time-frames to identify new lead agents; there is an urgent need for better pre-clinical decision making stages to understand targets, lead selection, and late-stage attribution. Until recently, the search for drug targets has focused on relatively small parts of the regulatory network under the assumption that key events can be controlled by targeting single pathways. Since it is now becoming clear that these early assumptions may not hold and successful treatments are likely to employ drugs that simultaneously target a number of different sites in the regulatory network, it is timely to redress this imbalance with mathematical models that represents complex biological regulatory system. Such models may help to identify feedback mechanisms that offset the effects of drugs and predict action as well non-desirable side effects.

7.6 Cellular informatics - screening approaches for drug discovery

High-content screening (HCS) is a drug discovery approach that combines modern cell biology with automated high-resolution microscopy and robotic handling (Harrison 2008). HCS describes the use of spatially or temporally resolved methods to discover more in an individual experiment than one single experimental value. HCS allows functional analysis of targets and pathway modulation in cells by drug compounds in a high content manner. HCS has emerged as a promising solution to improve the quality of decision making in drug discovery and development (Liptrot 2001). However, tools required for processing and analyzing HCS data are rather immature but allow read outs to assess parameters such as cytotoxicity, apoptosis, and effects on cell cycle (Crouch and Slater 2001; Liptrot 2001; Slater 2001). As a whole it provides an important link between molecular screening and functional cellular assays. A number of commercial software like MetaMorph from Molecular Devices Corporation (Sunnyvale, California) and open source software like CellProfiler (Carpenter et al. 2006) been developed and used for quantifying and analyzing cellular behaviour in HCS fashion.

7.6.1 Integrating static-dynamic and cross platform data

High-through-put (HTS) and indeed HCS, incorporating elegant reporter assays, have been effectively used to profile drugs based on simple stimulus-response readouts, however the design of current high-content instrumentation, discards biological heterogeneity and most assays never contend with dynamic processes (White and Errington 2005). In the absence of detailed kinetic information, simple snap-shot or static high-content-assays that measure drug effects provide an over simplified and often skewed view of the nature of resistant and sensitive cells. ProgeniTRAK and FluorTRAK address this need and output the lineage map that encapsulate kinetic data derived from image sources and consequently permit a coherent analysis of drug-induced perturbations in complex heterogeneous populations. Integrating these kinetics measurements with end-point results will generate a comprehensive view on drug effect on cellular behaviour. From a data volume perspective, arguably the kinetic measurements even with the automated tracking algorithms will never match the static HCS assays, but integration will establish the “missing link” between two successive static points and thus will enable us to understand the dynamics processes that lead from one static point to another.

Again the rich '-omic' data generated from different sources (e.g. Entrez, EBI etc.) are snap shot of the dynamic process and setting these data in cellular context requires lineage map which provides the opportunity to integrate genomic and or proteomic data and subsequently facilitates downward understanding – from cell to molecule. For example, integrating cyclin B1 expression profile encoded through FluorTRAK with cyclin B1 gene expression profile measured through microarray technology, would provide a comprehensive knowledge about cyclin B1 in a dynamic context thus augmenting our present understanding about the role of cyclin B1 in cell cycle. Thus the lineage map provides the perspective to integrate information from downwards ('-omic' information) and upwards to the organism (see figure 1-1). Sydney Brenner (Nobel laureate, 2002) in his Nobel Laureate speech stressed the importance of integration of data at this cellular level – "meso scale" for abstraction of knowledge about the dynamic process of life; and Mark Ellisman (University of Washington) in his recent (April 2008) speech on 7th annual symposium on Systems Biology and Engineering; scaled this meso scale within 1 nM to 100 μM range within which different cellular components (synaptic cleft to whole cell) can be visualized and quantified through imaging technology.

7.7 Future perspectives - scaling up of the infrastructure to a community level resource

HCS and HTS are widely used in both a systems biology and drug discovery context as these technologies describe cell phenotypes which enable broad, quantitative and machine readable measures of the responses of cell population to perturbation (Eisen et al. 1998; Gavin et al. 2002; Ho et al. 2002; Uetz et al. 2000) analysis of the temporal and spatial changes in cells and cell constituents in cellular arrays (Palmer and Freeman 2005) has the potential to create enormous systems biology knowledge bases. Moreover HCS is also been employed along with a range of early drug discovery platforms, including lead optimization where new knowledge is being used to facilitate the decision – making process (Giuliano et al. 2005). It is clear that for cell-based analysis to keep pace with other HTS oriented applications such as microarray technologies and proteomics, significant inroads into toolbox development must take place. Through this infrastructure proving ground we have developed a strategy for overcoming the current lack of informatics frameworks in microscopy and image analysis, we focused not on the hardware solutions but on the embedded and linked informatics and minimal standards required for encoding cellular data into a lineage map format that encapsulates the multi-scalar kinetic information of a living cell. This 'proof-of-concept' infrastructure and associated validation studies have indicated the effectiveness of a lineage map data format approach and for the first time have provided the scope to interrogate the complex interplay of cellular dynamics in different biological processes (eg cell cycle, stem cell biology and ageing). In addition these encoded lineages have facilitated a metadata level understanding, bridging the gap between biologists and mathematicians.

This current work has allowed us to identify the critical steps and bottlenecks required to scale-up such a prototype infrastructure to contemplate other HCS applications in this context. The manual image analysis approach adopted by ProgeniTRAK/FluorTRAK is indeed time consuming and tedious work although it allows for precise data encoding, particularly resolving the outcome of mitosis such as identifying abnormal outcomes (cells undergoing furrow regression and entering polyploidy). Therefore semi- to full-automation of cell tracking and lineage construction would enhance the through-put for data processing; this remains the current bottleneck for microscopy based cell assays. Phase-contrast transmission microscopy offers a probeless and non-perturbing contrast mode, providing low resolution but highly informative outputs on cell behaviour (e.g. cell shape and cell position). The changes in these two basic features facilitate assays describing critical global morphological cell responses such as cell division, cell death and cell motility (White and Errington 2005). The automation of cell tracking raises many challenges - the combination of low signal-to-noise ratio of phase contrast microscopy images, high and varying densities of the cell cultures, topological complexities of cell shapes, and wide range of cell behaviour poses many challenges to existing tracking techniques. With the recent advancement in tracking algorithms that addresses the principal challenges of single particle tracking (Li et al. 2008) it is evident that transforming of such prototype infrastructure to the HTS domain is indeed a reality.

However to achieve such ambitions, these algorithms must possess the ability identify each of the node information making up the lineage, the minimal information therefore is the ability to the cell division event and the associated outcomes, in addition a second event called cell death must also be identified. Therefore a self checking algorithm will also need to be developed to ensure the cells are tracked correctly, this will involve cross checking images tracked forward and backwards in time e.g. in reverse time sequence where the two daughter cells 'merge' to form the parent cell. In the initial stages further information such as frame-to-frame cell morphology could be considered as second level information; therefore removing the requirement of a cell edge segmentation algorithm thus alleviating the demands of tracking algorithms. This immediate raises the issue of the minimal standard associated with the lineage map.

Resolving tracking through-put will only put the HTS leverage to the infrastructure, and to scale it to a community based infrastructure, a standardised data format remains the critical focus, as sharing data requires adherence to standards (data format and semantics) and protocols (for access and exchange). Thus a requirement for a minimal standard needs to be modelled and validated, similar to the minimal information standards that have been used for microarray data (Brazma et al. 2001) and recently for flow cytometry data (Lee et al. 2008), the latter now reaching out to an international community under the auspices of ISAC (International Society for Analytical Cytometry). A proposed lineage minimal standard needs to interface with other minimal information standards such as MIFlowCyt (Minimal information about Flow Cytometry Experiments) (Lee et al. 2008); MIACA (Minimal Information About Cellular Assays); MIBBI (Minimum reporting guidelines for biological and biomedical investigations) (Taylor et al. 2008) and OME (Open Microscopy Environment) (Swedlow et al. 2003). Therefore a suggested minimal standard called - a Minimal Information about Cell Lineages And Dynamics (MICLAD) would build and interlink with the existing standards in cell-based analysis. MICLAD will be a hybrid data model that encompasses experimental, morphological and behavioural attributes of a cell in a lineage format and thus enabling organizing, analyzing, interpreting and sharing cellular dynamics data within such infrastructure. It is envisaged that such data format will utilize XML which facilitate both the data exchange and future extension of the data format. Such approach will facilitate data exchange not only within the users of such infrastructure but also with OME/CellProfiler/MIFlowCyt users, moreover will pave the foundation of interacting with mathematical modellers who utilize current XML based cell modelling and simulation approaches like Systems Biology Workbench (Sauro et al. 2003), CellML (Cuellar et al. 2003). Thus in combination with the cell tracking algorithms a standardized data format will transform the prototype infrastructure presented in this thesis to a level which not only will resonate with other HTS/HCS technologies, but also will provide the framework for a cohesive data sharing across the life sciences, engineering and mathematics disciplines alike.

7.8 Concluding remarks

Transforming images to knowledge, the central dogma introduced and demonstrated through this bioinformatics infrastructure, has provided a new level of understanding about cellular processes. More importantly this has invoked another level of data perception as well as providing tools for complex hypothesis testing. At the basic level the bioinformatics framework developed in this thesis has provided the integrative environment encompassing the Modus Operandi of linking the biologist with the mathematicians. This thesis however goes further and addresses the concept that the cell lineage map represents the minimal unit for defining a dynamic cell system, reaching down to molecular networks while at the same time sustaining information that defines cellular relationships and interactions necessary for maintaining a multi-cellular community.

Appendix

All Appendices are attached with the Compact Disk (CD) provided. Please click to the "Appendix.doc" file in the CD to reveal the list of appendices including the electronic version of this PhD thesis.

Bellow is an outline of the table of content for the Appendix CD.

Appendix No	Description	File Link
Appendix IV	Lists of attributes encoded by ProgeniTRAK	X
Appendix V	ProgeniTRAK manual for encoding	W
Appendix VI	ProgeniTRAK source code	★
Appendix VII	Installation manual for ProgeniTRAK	W
Appendix VIII	Drug screening protocol	W
Appendix IX	Timelapse image sequence showing unperturbed growth of U-2 OS cell line in a 112 h experiment	AVI
Appendix X	Johnson curve fit on IMT distribution. Summer masters project done by Chris Hedley (Summer 2005)	W
Appendix XI	Protocol describing Fluorescence timelapse experiment.	W
Appendix XII	Fluorescence timelapse image sequence	AVI
Appendix XIII	FluroTRAK Source code	★
Appendix XIV	Lists of attributes encoded by FluroTRAK	X
Appendix XV	Encoding manual for FluroTRAK	W
Appendix XVI	Installation manual for intensity viewer	W
Appendix XVII	Mathematical model derived from BrdUrd pulse-and-chase analysis.	W
Appendix XVIII	Methods of counting cells from coulter counter	W

	Related Publications	
Appendix XIX	Smith PJ, Chin SF, Njoh K, **Khan IA**, Chappell MJ, Errington RJ (2008) Cell cycle checkpoint-guarded routes to catenation-induced chromosomal instability. ***SEB Exp Biol Ser.*** 2008;59:219-42. PMID: 18368926	e
Appendix XX	Judith Perez-Velazquez*, Neil D. Evans, Michael Chappell, Rachel Errington, Paul Smith, **Imtiaz Khan** (*2008*) A Mathematical Model of Cyclin B1 Dynamics at the Single Cell Level in Osteosarcoma Cells. Accepted paper for ***17th IFAC World Congress***, Seoul, Korea.	PDF
Appendix XXI	Chappell MJ, Evans ND, Errington RJ, **Khan IA**, Campbell L, Ali R, Godfrey KR, Smith PJ (*2008*) A coupled drug kinetics-cell cycle model to analyse the response of human cells to intervention by topotecan. ***Comput Methods Programs Biomed***. 89(2):169-78. PMID: 18082908	PDF
Appendix XXII	Paul J Smith, Nuria Marquez, Marie Wiltshire, Sally Chappell, Kerenza Njoh, Lee Campbell**, Imtiaz A Khan**, Oscar R Silvestre and Rachel J Errington (*2007*) Mitotic Bypass Via An Occult Cell Cycle Phase Following DNA Topoisomerase II Inhibition In p53 Functional Human Tumour Cells, ***Cell Cycle***, 6:16, 2071-81 [PMID: 17721081]	PDF
Appendix XXIII	**I.A. Khan**, J. Fisher, P. J. Smith and R. J. Errington (*2007*) A Bioinformatics approach for the interrogation of molecular events in single cells: transforming fluorescent timelapse microscopy images into numbers. ***BMC Systems Biology*** (2007) 1(Suppl 1):P31	PDF
Appendix XXIV	**I.A. Khan**, P. Husemann, L. Campbell, N. S. White, R. White, P. J. Smith and R.J. Errington (*2007*) ProgeniDB: A novel cell lineage database for generation associated phenotypic behavior in cell-based assays ***Cell Cycle***, 6:7, 868-74 [PMID: 17387278].	PDF
Appendix XXV	Paul J Smith, **I.A Khan** and R.J Errington (*2007*) Cytomics And Drug Development. ***Cytometry A***. 71A:349–351. [PMID: 17323350]	PDF
Appendix XXVI	***Khan IA*, **Hedley CJ, White NS, Ali R, Chappell MJ, Evans ND, Campbell L, Marquez N, Fisher J, Smith PJ, Errington RJ. (*2006*) A novel integrative bioinformatics environment for encoding and interrogating timelapse microscopy images. In: Feng, DD ed. Modelling and Control in Biomedical Systems, Reims, France. Amsterdam, ***ELSEVIER***, pp 273-8	PDF

References

Abal, M. et al. 2003. Taxanes: microtubule and centrosome targets, and cell cycle dependent mechanisms of action. *Curr Cancer Drug Targets* 3(3), pp. 193-203.

Abbott, D. et al. eds. 1998. *IEEE International Conference on Systems, Man, and Cybernetics* San Diego, CA:

Abraham, V. C. et al. 2004. High content screening applied to large-scale cell biology. *Trends Biotechnol* 22(1), pp. 15-22.

Abramowitz, M. and Davidson, M. 2007. Introduction to Microscopy.*Molecular Expressions*.

Abrous, D. N. et al. 2005. Adult neurogenesis: from precursors to network and physiology. *Physiol Rev* 85(2), pp. 523-569.

Agard, D. A. et al. 1989. Fluorescence microscopy in three dimensions.*Methods in Cell Biology*. Vol. 30. New York: Academic, pp. 353-377.

Aguda, B. D. 1999. A quantitative analysis of the kinetics of the G(2) DNA damage checkpoint system. *Proc Natl Acad Sci U S A* 96(20), pp. 11352-11357.

Aguda, B. D. and Tang, Y. 1999. The kinetic origins of the restriction point in the mammalian cell cycle. *Cell Prolif* 32(5), pp. 321-335.

Alarcon, T. et al. 2004. A mathematical model of the effects of hypoxia on the cell-cycle of normal and cancer cells. *J Theor Biol* 229(3), pp. 395-411.

Alarcon, T. et al. 2006. Mathematical models of the fate of lymphoma B cells after antigen receptor ligation with specific antibodies. *J Theor Biol* 240(1), pp. 54-71.

Alarcon, T. and Tindall, M. J. 2007. Modelling cell growth and its modulation of the G1/S transition. *Bull Math Biol* 69(1), pp. 197-214.

Alfieri, R. et al. 2007. A data integration approach for cell cycle analysis oriented to model simulation in systems biology. *BMC Syst Biol* 1, p. 35.

Allen, F. et al. 2001. Blue Gene: A vision for protein science using a petaflop supercomputer. *IBM SYSTEMS JOURNAL* 40(2).

Allman, R. et al. 2003. Delayed expression of apoptosis in human lymphoma cells undergoing low-dose taxol-induced mitotic stress. *Br J Cancer* 88(10), pp. 1649-1658.

Aloy, P. et al. 2003. Predictions without templates: new folds, secondary structure, and contacts in CASP5. *Proteins* 53, pp. 436-456.

Altman, D. 1991. Practical Statistics for Medical Research. London: Chapman & Hall, pp. 285-288.

Altschul, S. F. et al. 1990. Basic local alignment search tool. *J Mol Biol* 215(3), pp. 403-410.

Alvarez-Buylla, A. et al. 2001. A unified hypothesis on the lineage of neural stem cells. *Nat Rev Neurosci* 2(4), pp. 287-293.

Anderson, D. J. et al. 2001. Can stem cells cross lineage boundaries? *Nat Med* 7(4), pp. 393-395.

Ardavin, C. et al. 2001. Origin and differentiation of dendritic cells. *Trends Immunol* 22(12), pp. 691-700.

Arkhipov, S. N. et al. 2005. Chemical cytometry for monitoring metabolism of a Rasmimicking substrate in single cells. *Cytometry Part A* 63(A), p. 41.

Arnaud, L. et al. 1998. GFP tagging reveals human Polo-like kinase 1 at the kinetochore/centromere region of mitotic chromosomes. *Chromosoma* 107(6-7), pp. 424-429.

Arndt-Jovin, D. J. et al. 1985. Fluorescence digital imaging microscopy in cell biology. *Science* 230(4723), pp. 247-256.

Axelrod, D. 1989a. Fluorescence Microscopy of Living Cells in Culture, Part B: Quantitative Fluorescence Microscopy -Imaging and Spectroscopy.*Methods in Cell Biology*. Vol. 30. New York: Academic.

Axelrod, D. 1989b. Fluorescence polarization microscop.*Methods in Cell Biology*. Vol. 30. New York: Academics, pp. 333-352.

Axelrod, D. 1989c. Total internal reflection fluorescence microscopy.*Methods in Cell Biology*. Vol. 30. New York: Academic, pp. 246-270.

Bai, C. et al. 1994. Human cyclin F. *The EMBO Journal* 13(24), pp. 6087-6098.

Bajer, A. S. and Bajer, J. M. 1972. Spindle dynamics and chromosome movements. *International Review of Cytology* supp. 3, pp. 1-271.

Bernards, R. and Weinberg, R. A. 2002. A progression puzzle. *Nature* 418(6900), p. 823.

Bertuzzi, A. et al. 1988. Constrained cross-validation applied to estimation of kinetic parameters of cell populations in perturbed growth. In: Iri, M. and Yajima, K. eds. *System Modelling and Optimization*. Berlin: Springer, pp. 614-623.

Bocsi, J. et al. 2004a. Related scanning fluorescent microscopy analysis is applicable for absolute and relative cell frequency determinations. *Cytometry Part A* 61(A), pp. 1-4.

Bocsi, J. et al. 2004b. Scanning fluorescent microscopy analysis is applicable for absolute and relative cell frequency determinations. *Cytometry A* 61(1), pp. 1-8.

Bou-Gharios, G. et al. 2004. Extra-cellular matrix in vascular networks. *Cell Prolif* 37(3), pp. 207-220.

Brakenhoff, G. J. et al. 1990. Potentialities and limitations of confocal microscopy for the study of 3-dimensional biological structures. In: Herman, B. and Jacobson, K. eds. *Optical Microscopy for Biology*. New York: Wiley, pp. 19-28.

Braun, V. et al. 2003. ALES: cell lineage analysis and mapping of developmental events. *Bioinformatics* 19(7), pp. 851-858.

Bray, D. 2003. Molecular networks: the top-down view. *Science* 301(5641), pp. 1864-1865.

Brazma, A. et al. 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 29(4), pp. 365-371.

Brenner, S. 2001. General Discussion II. Novartis Foundation Symposium. 239(67), pp. 150-159.

Brenner, S. ed. 2003. *CHI Conference* Santa Clara, CA:

Bright, G. R. et al. 1989. Fluorescence ratio imaging microscopy.*Methods in Cell Biology*. Vol. 30. New York: Academic, pp. 157-192.

Bruggeman, F. J. et al. 2007. Introduction to systems biology. *Exs* 97, pp. 1-19.

Bullen, A. 2008. Microscopic imaging techniques for drug discovery. *Nat Rev Drug Discov* 7(1), pp. 54-67.

Bunyak, F. et al. eds. 2006. *3rd IEEE International Symposium on Biomedical Imaging: From Nano to Macro* Arlington, VA:

Burley, S. K. 2000. An overview of structural genomics. *Nat Struct Biol* 7 Suppl, pp. 932-934.

Cardullo, R. A. and Parpura, V. 2003. Fluorescence resonance energy transfer microscopy: theory and instrumentation. *Methods in Cell Biology* 72, pp. 415-430.

Carnero, A. 2002. Targeting the cell cycle for cancer therapy. *Br J Cancer* 87(2), pp. 129-133.

Carpenter, A. E. et al. 2006. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol* 7(10), p. R100.

Casasnovas, R. O. et al. 2003. Immunological classification of acute myeloblastic leukemias: relevance to patient outcome. *Leukemia* 17, pp. 515-517.

Castells, M. 2001. Lessons from the History of Internet.*The Internet Galaxy*. Oxford Univ. Press., pp. 9-35.

Cavanna, T. et al. 2007. Evidence for protein 4.1B acting as a metastasis suppressor. *J Cell Sci* 120(Pt 4), pp. 606-616.

Cervinka, M. et al. 2008. The role of time-lapse fluorescent microscopy in the characterization of toxic effects in cell populations cultivated in vitro. *Toxicol In Vitro*.

Chang, F. and Drubin, D. G. 1996. Cell division: why daughters cannot be like their mothers. *Curr Biol* 6(6), pp. 651-654.

Chappell, M. J. et al. 2008. A coupled drug kinetics-cell cycle model to analyse the response of human cells to intervention by topotecan. *Comput Methods Programs Biomed* 89(2), pp. 169-178.

Charrier-Savournin, F. B. et al. 2004. p21-Mediated nuclear retention of cyclin B1-Cdk1 in response to genotoxic stress. *Mol Biol Cell* 15(9), pp. 3965-3976.

Chen, K. C. et al. 2000. Kinetic analysis of a molecular model of the budding yeast cell cycle. *Mol Biol Cell* 11(1), pp. 369-391.

Chisholm, A. D. and Hodgkin, J. 1989. The mab-9 gene controls the fate of B, the major male-specific blast cell in the tail region of Caenorhabditis elegans. *Genes Dev* 3(9), pp. 1413-1423.

Chu, K. et al. 2002. Computerized Video Time-Lapse (CVTL) Analysis of Cell Death Kinetics in Human Bladder Carcinoma Cells (EJ30) X-Irradiated in Different Phases of the Cell Cycle. *Rad. Res.* 158, pp. 667-677.

Chu, K. et al. 2004. Computerized Video Time Lapse Study of Cell Cycle Delay and Arrest, Mitotic Catastrophe, Apoptosis and Clonogenic Survival in Irradiated 14-3-3s and CDKN1A (p21) Knockout Cell Lines. *Rad. Res.* 162, pp. 270-286.

Clarke, J. D. and Tickle, C. 1999. Fate maps old and new. *Nat Cell Biol* 1(4), pp. E103-109.

Cliby, W. A. et al. 2002. S phase and G2 arrests induced by topoisomerase I poisons are dependent on ATR kinase function. *J Biol Chem* 277(2), pp. 1599-1606.

Clyde, R. et al. 2006. The role of modelling in identifying drug targets for diseases of the cell cycle *J. R. Soc. Interface* 10, pp. 617-627.

Collins, T. J. 2007. ImageJ for microscopy. *Biotechniques* 43(1 Suppl), pp. 25-30.

Conrad, C. et al. 2004. Automatic identification of subcellular phenotypes on human cell arrays. *Genome Res* 14(6), pp. 1130-1136.

Crouch, S. and Slater, K. 2001. High-throughput cytotoxicity screening: hit and miss. *Drug Discov. Today* 6(48-531).

Csika´sz-Nagy, A. et al. 2006. Analysis of a Generic Model of Eukaryotic Cell-Cycle Regulation. *Biophysical Journal* 90, pp. 4361-4379.

Cuellar, A. et al. 2003. CellML 1.1 for the Definition and Exchange of Biological Models. *Conf. Proc. IFAC Symposium on Modelling and Control in Biomedical Systems*, pp. 451-456.

Dawkins, R. 1976. *The Selfish Gene.* Oxford and New York: Oxford Univ. Press.

Dayhoff, M. 1979. *Atlas of Protein Sequence and Structure.*

Demidenko, Z. N. et al. 2008. Mechanism of G1-like arrest by low concentrations of paclitaxel: next cell cycle p53-dependent arrest with sub G1 DNA content mediated by prolonged mitosis. *Oncogene.*

Devault, A. et al. 1991. Concerted roles of cyclin A, cdc25+ mitotic inducer, and type 2A phosphatase in activating the cyclin B/cdc2 protein kinase at the G2/M phase transition. *Cold Spring Harbor Symposia on Quantitative Biology* 56, pp. 503-513.

Dor, Y. et al. 2004. Adult pancreatic beta-cells are formed by self-duplication rather than stem-cell differentiation. *Nature* 429(6987), pp. 41-46.

Dove, A. 1999. Proteomics: translating genomics into products? *Nat Biotechnol* 17(3), pp. 233-236.

Dovichi, N. J. and Hu, S. 2003. Chemical cytometry. *Curr Opin Chem Biol* 7, p. 603.

Dufour, A. et al. 2005. *Segmenting and tracking fluorescent cells in dynamic 3-D microscopy with coupled active surfaces.* IEEE Trans. Image Process. pp. 1396–1410.

Dunn, G. A. et al. 2004. Fluorescence localization after photobleaching (FLAP). *Curr Protoc Cell Biol* Chapter 21, p. Unit 21 22.

Echeverri, C. J. and Perrimon, N. 2006. High-throughput RNAi screening in cultured cells: a user's guide. *Nat Rev Genet* 7(5), pp. 373-384.

Ecker, R. C. et al. 2004a. Microscopy-based multicolor tissue cytometry at the single cell level. *Cytometry Part A* 59(A), pp. 182-184.

Ecker, R. C. et al. 2004b. Application of spectral imaging microscopy in cytomics and fluorescence eneregy transfer (FRET) analysis. *Cytometry Part A* 59(A), pp. 172-174.

Ecker, R. C. and Steiner, G. E. 2004. Microscopy-based multicolor tissue cytometry at the single-cell level. *Cytometry A* 59(2), pp. 182-190.

Ecker, R. C. and Tarnok, A. 2005. Cytomics goes 3D: towards cytomics. *Cytometry Part A* 65(A), pp. 1-3.

Edwards, B. S. et al. 2004. Flow cytometry for high-throughput, high-content screening. *Curr Opin Chem Biol* 8, p. 392.

Eggert, U. S. and Mitchison, T. J. 2006. Small molecule screening by imaging. *Curr Opin Chem Biol* 10(3), pp. 232-237.

Eisen, M. B. et al. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95(25), pp. 14863-14868.

Endlich, B. et al. 2000. Computerized Video Time-Lapse Microscopy Studies of Ionizing Radiation-Induced Rapid-Interphase and Mitosis-Related Apoptosis in Lymphoid Cells. *Rad. Res.* 153, pp. 36-48.

Erickson, D. 2003. Wanted: drug hunters *in vivo*. *The Business and Medicine Report* 21, pp. 45-52.

Errington, R. et al. 2006. Time□lapse microscopy approaches to track cell cycle progression at the single cell.*Current Protocols in Cytometry*. New Jersey: J Wiley.

Farkas, D. L. et al. 1993. Multimode light microscopy and the dynamics of molecules, cells, and tissues. *Annu Rev Physiol* 55, pp. 785-817.

Farooqui, R. and Fenteany, G. 2005. Multiple rows of cells behind an epithelial wound edge extend cryptic lamellipodia to collectively drive cell-sheet movement. *J. Cell Science* 118(1), pp. 51-63.

Feeney, G. P. et al. 2003. Tracking the cell cycle origins for escape from topotecan action by breast cancer cells. *Br J Cancer* 88(8), pp. 1310-1317.

Fidler, I. J. and Hart, I. R. 1982. Biological diversity in metastatic neoplasms: origins and implications. *Science* 217(4564), pp. 998-1003.

Figeys, D. 2004. Combining different 'omics' technologies to map and validate protein-protein interactions in humans. *Brief Funct Genomic Proteomic* 2(4), pp. 357-365.

Fitch, D. H. and Emmons, S. W. 1995. Variable cell positions and cell contacts underlie morphological evolution of the rays in the male tails of nematodes related to Caenorhabditis elegans. *Dev Biol* 170(2), pp. 564-582.

Forrester, H. et al. 2000. Computerized Video Time-Lapse Analysis of Apoptosis of REC:Myc Cells X-Irradiated in Different Phases of the Cell Cycle. *Rad. Res.* 154, pp. 625-639.

Forrester, H. B. et al. 1999. Using computerized video time lapse for quantifying cell death of X-irradiated rat embryo cells transfected with c-myc or c-Ha-ras. *Cancer Res* 59(4), pp. 931-939.

Foster, I. 2005. Service-oriented science. *Science* 308(5723), pp. 814-817.

Fox, G. et al. 2003. Grid Computing: Making the Global Infrastructure a Reality. In: Berman, F. et al. eds. *Overview of grid computing environments.* Chichester: Wiley.

Frumkin, D. et al. 2005. Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput Biol* 1(5), p. e50.

Fung, T. K. and Poon, R. Y. 2005. A roller coaster ride with the mitotic cyclins. *Seminars Cell Developmental Biology* 16(3), pp. 335-342.

Galperin, M. Y. 2008. The Molecular Biology Database Collection: 2008 update. *Nucleic Acids Res* 36(Database issue), pp. D2-4.

Gavin, A. C. et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415(6868), pp. 141-147.

Gaziova, I. and Bhat, K. M. 2007. Generating asymmetry: with and without self-renewal. *Prog Mol Subcell Biol* 45, pp. 143-178.

George, T. C. et al. 2004. Distingusihing modes of cell death using the ImageStream multispectral imaging flow cytometer. *Cytometry Part A* 59(A), p. 237.

Gerstner, A. O. et al. 2006. Comparison of immunophenotyping by slide-based cytometry and by flow cytometry. *J Immunol Methods* 311(1-2), pp. 130-138.

Gerstner, A. O. et al. 2004. Quantitative histology by multicolor slide-based cytometry. *Cytometry A* 59(2), pp. 210-219.

Giuliano, K. A. 2003. High-content profiling of drug-drug interactions: cellular targets involved in the modulation of microtubule drug action by the antifungal ketoconazole. *J Biomol Screen* 8(2), pp. 125-135.

Giuliano, K. A. et al. 2004. High-content screening with siRNA optimizes a cell biological approach to drug discovery: defining the role of P53 activation in the cellular response to anticancer drugs. *J Biomol Screen* 9(7), pp. 557-568.

Giuliano, K. A. et al. 2005. Systems cell biology knowledge created from high content screening. *Assay Drug Dev Technol* 3(5), pp. 501-514.

Giuliano, K. A. et al. 1997. High-content screening: A new approach to easing key bottlenecks in the drug discovery process. *Journal of Biomolecular Screening* 2(4), pp. 249-259.

Goldberg, I. G. et al. 2005. The Open Microscopy Environment (OME) Data Model and XML file: open tools for informatics and quantitative analysis in biological imaging. *Genome Biol* 6(5), p. R47.

Goldbeter, A. 1991. A minimal cascade model for the mitotic oscillator involving cyclin and cdc2 kinase. *Proc Natl Acad Sci U S A* 88(20), pp. 9107-9111.

Goldstein, D. J. 1982. A simple quantitative analysis of phase contrast microscopy, not restricted to objects of very low retardation. *J Microsc* 128(Pt 1), pp. 33-47.

Guo, S. and Kemphues, K. J. 1996. Molecular genetics of asymmetric cleavage in the early Caenorhabditis elegans embryo. *Curr Opin Genet Dev* 6(4), pp. 408-415.

Hagen, J. B. 2000. The origins of bioinformatics. *Nat Rev Genet* 1(3), pp. 231-236.

Haraguchi, T. 2002. Live cell imaging: approaches for studying protein dynamics in living cells. *Cell Struct Funct* 27(5), pp. 333-334.

Harrigan, G. and Goodacre, R. 2003. *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*. Boston: Kluwer Academic Publishers.

Harrison, C. 2008. High-content screening: Integrating information. *Nature Reviews Drug Discovery* 7, pp. 121-121.

Hastings, S. et al. 1977. Existence of periodic solutions for negative feedback cellular control systems. *J. Differ. Equations.* 25, pp. 39-64.

Haugland, R. P. 1992. *Handbook of Fluorescent Probes and Research Chemicals.* . Eugene.

Heidorn, P. B. et al. 2007. Biological information specialists for biological informatics. *J Biomed Discov Collab* 2, p. 1.

Henikoff, S. and Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89(22), pp. 10915-10919.

Hey, T. and Trefethen, A. 2003. e-Science and its implications. *Philos Transact A Math Phys Eng Sci* 361(1809), pp. 1809-1825.

HGP, U. D. o. E. G. R. 2003. Genomics and Its Impact on Science and Society *The Human Genome Project and Beyond.*

Higashikubo, R. et al. 1996. Flow cytometric BrdUrd-pulse-chase study of X-ray-induced alterations in cell cycle progression. *Cell Prolif* 29(1), pp. 43-57.

Ho, Y. et al. 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 415(6868), pp. 180-183.

Hope, K. J. et al. 2004. Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. *Nat Immunol* 5(7), pp. 738-743.

Horvitz, H. R. and Herskowitz, I. 1992. Mechanisms of asymmetric cell division: two Bs or not two Bs, that is the question. *Cell* 68(2), pp. 237-255.

Huang, J. and Raff, J. W. 1999. The disappearance of cyclin B at the end of mitosis is regulated spatially in Drosophila cells. *Embo J* 18(8), pp. 2184-2195.

Huang, X. et al. 2003. DNA damage induced by DNA topoisomerase I- and topoisomerase II-inhibitors detected by histone H2AX phosphorylation in relation to the cell cycle phase and apoptosis. *Cell Cycle* 2(6), pp. 614-619.

Inoue, S. 1989. Imaging of unresolved objects, superresolution, and precision of distance measurement with video microscopy. In: Taylor, D. and Wang, Y. eds. *Methods in Cell Biology*. Vol. 30. New York: Academic, pp. 85-112.

Inoue, S. 1990. Whither video microscopy? Towards 4-D imaging at the highest resolution of the light microscope. In: Herman, B. and Jacobson, K. eds. *Optical Microscopy for Biology*. New York: Wiley-Liss, pp. 497-511.

Inoué, S. 1986. *Video Microscopy.* 1st ed. New York: Plenum Press.

Inoué, S. and Spring, K. 1997. *Video Microscopy.* 2nd ed. New York: Plenum Press.

Ivan, L. C. and Greulich, R. C. 1963. Evidence for an essentially constant duration of DNA synthesis in renewing epithelia of the adult mouse. *The Journal of Cell Biology* 18, pp. 31-40.

Jan, Y. N. and Jan, L. Y. 1998. Asymmetric cell division. *Nature* 392(6678), pp. 775-778.

Jensen, O. N. 2006. Interpreting the protein language using proteomics. *Nat Rev Mol Cell Biol* 7(6), pp. 391-403.

Johnson, N. 1949. Systems of frequency curves generated by methods of translation. *Biometrika* 36, pp. 149-176.

Johnson, S. 1967. Hierarchical Clustering Schemes. *Psychometrika* 2, pp. 241-254.

Jonsson, P. F. et al. 2006. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics* 7, p. 2.

Jovcic, G. et al. 2004. In vivo effects of interleukin-17 on haematopoietic cells and cytokine release in normal mice. *Cell Prolif* 37(6), pp. 401-412.

Jovin, T. M. and Amdt-Jovin, D. J. 1989. Luminescence digital imaging microscopy. *Annual Review of Biophysics and Biophysical Chemistry* 18, pp. 271-308.

Joyce, A. R. and Palsson, B. O. 2006. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* 7(3), pp. 198-210.

Kanda, T. et al. 1998. Histone-GFP fusion protein enables sensitive analysis of chromosome dynamics in living mammalian cells. *Curr Biol* 8(7), pp. 377-385.

Kanehisa, H. 2000. *Post-Genome Informatics.* Oxford: Oxford Univ. Press.

Kanehisa, H. and Brok, P. 2003. Bioinformatics in the post-sequence era. *Nat Genet* 33(Supp), pp. 305-310.

Kanehisa, M. 2002. The KEGG database. *Novartis Found Symp* 247, pp. 91-101; discussion 101-103, 119-128, 244-152.

Kantor, A. B. et al. 2004. Immune systems biology: immunoprofiling of cell and molecules. *BioTechniques* 36, p. 520.

Karp, G. 2007. *Cell and Molecular Biology: Concepts and Experiments.* 5 ed. New York: Wiley.

Kerns, E. H. et al. 2003. Pharmaceutical profiling method for lipophilicity and integrity using liquid chromatography-mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci* 791(1-2), pp. 381-388.

Khan, I. A. et al. 2007. ProgeniDB: a novel cell lineage database for generation associated phenotypic behavior in cell-based assays. *Cell Cycle* 6(7), pp. 868-874.

Kim, K. M. and Shibata, D. 2002. Methylation reveals a niche: stem cell succession in human colon crypts. *Oncogene* 21(35), pp. 5441-5449.

King, K. L. and Cidlowski, J. A. 1998. Cell cycle regulation and apoptosis. *Annu Rev Physiol* 60, pp. 601-617.

Knowles, J. and Gromo, G. 2003. A guide to drug discovery: Target selection in drug discovery. *Nat Rev Drug Discov* 2(1), pp. 63-69.

Kollmannsberger, C. et al. 1999. Topotecan - A novel topoisomerase I inhibitor: pharmacology and clinical experience. *Oncology* 56(1), pp. 1-12.

Kong, M. et al. 2000. Cyclin F regulates the nuclear localization of cyclin B1 through a cyclin-cyclin interaction. *The EMBO Journal* 19(6), pp. 1378-1388.

Kraut, R. et al. 1996. Role of inscuteable in orienting asymmetric cell divisions in Drosophila. *Nature* 383(6595), pp. 50-55.

Kriete, A. 2005. Cytomics in the realm of systems biology. *Cytometry A* 68(1), pp. 19-20.

Kriete, A. and Boyce, K. 2005. Automated tissue analysis - a bioinformatics perspective. *Methods Inf Med* 44, pp. 32-35.

Kuczek, T. and Axelrod, D. E. 1987. Tumor cell heterogeneity: divided-colony assay for measuring drug response. *Proc Natl Acad Sci U S A* 84(13), pp. 4490-4494.

Kushner, J. et al. 1999. Aberrant expression of cyclin A and cyclin B1 proteins in oral carcinoma. *J Oral Pathol Med* 28(2), pp. 77-81.

Lang, P. et al. 2006. Cellular imaging in drug discovery. *Nat Rev Drug Discov* 5(4), pp. 343-356.

Lee, J. A. et al. 2008. MIFlowCyt: the minimum information about a Flow Cytometry Experiment. *Cytometry A* 73(10), pp. 926-930.

Lee, M. and Zaho, R. 2006. Functional Regulation of CIP/KIP CDK Inhibitors. In: Smith, P. and Yue, E. eds. *Inhibitors of Cyclin-dependent Kinases as Anti-tumor Agents.* Boca Raton, Florida: Taylor & Francis, pp. 29-55.

Levenson, R. M. et al. 2008. Multiplexing with multispectral imaging: from mice to microscopy. *Ilar J* 49(1), pp. 78-88.

Li, J. et al. 1997. Nuclear localization of cyclin B1 mediates its biological activity and is regulated by phosphorylation. *Proc Natl Acad Sci U S A* 94(2), pp. 502-507.

Li, K. et al. 2008. Cell population tracking and lineage construction with spatiotemporal context. *Med Image Anal* 12(5), pp. 546-566.

Lincoln, S. 2001. Genome annotation: from sequence to biology. *Nat. Rev. Genetics* 2, pp. 493-503

Lindqvist, A. et al. 2004. Characterisation of Cdc25B localisation and nuclear export during the cell cycle and in response to stress. *J Cell Sci* 117(Pt 21), pp. 4979-4990.

Lindsay, M. A. 2003. Target discovery. *Nat Rev Drug Discov* 2(10), pp. 831-838.

Lippincott-Schwartz, J. and Patterson, G. H. 2003. Development and use of fluorescent protein markers in living cells. *Science* 300, pp. 87-91.

Lippincott-Schwartz, J. et al. 2001. Studying protein dynamics in living cells. *Nature Reviews Molecular Cell Biology* 2(6), pp. 444-456.

Liptrot, C. 2001. High content screening - from cells to data to knowledge. *Drug Discov Today* 6(16), pp. 832-834.

Liu, W. and Chen, C. 2007. Cellular and multicellular form and function. *Adv Drug Deliv Rev.* 59, pp. 1319-1328.

Lockhart, D. J. and Winzeler, E. A. 2000. Genomics, gene expression and DNA arrays. *Nature* 405(6788), pp. 827-836.

Lodish, H. et al. 2004. *Molecular cell biology.* W.H.Freeman & Co Ltd.

Loew, L. M. e. 1988. *Spectroscopic Membrane Probes.* Boca Raton CRC Press, pp. Vol. I, 227 pp.; Vol. II, 206 pp.; Vot. III, 228 pp.

Loging, W. et al. 2007. High-throughput electronic biology: mining information for drug discovery. *Nat Rev Drug Discov* 6(3), pp. 220-230.

Lovell, M. J. and Mathur, A. 2004. The role of stem cells for treatment of cardiovascular disease. *Cell Prolif* 37(1), pp. 67-87.

Lupi, M. et al. 2004. Cytostatic and cytotoxic effects of topotecan decoded by a novel mathematical simulation approach. *Cancer Res* 64(8), pp. 2825-2832.

MacArthur, B. D. et al. 2006. A non-invasive method for in situ quantification of subpopulation behaviour in mixed cell culture. *J R Soc Interface* 3(6), pp. 63-69.

Maglott, D. et al. 2005. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 33(Database issue), pp. D54-58.

Malumbres, M. and Barbacid, M. 2001. To cycle or not to cycle: a critical decision in cancer. *Nat Rev Cancer* 1(3), pp. 222-231.

Marquez, N. et al. 2003. Single cell tracking reveals that Msh2 is a key component of an early-acting DNA damage-activated G2 checkpoint. *Oncogene* 22(48), pp. 7642-7648.

Marquez, N. et al. 2004. Microtubule stress modifies intra-nuclear location of Msh2 in mouse embryonic fibroblasts. *Cell Cycle* 3(5), pp. 662-671.

Maszewska, M. et al. 2002. Bromodeoxyuridine-labeled oligonucleotides as tools for oligonucleotide uptake studies. *Antisense Nucleic Acid Drug Dev* 12(6), pp. 379-391.

Mather, J. and Roberts, R. 1998. *Introduction to cell and tissue culture : theory and technique.* New York: Plenum Press.

Maynadié, M. et al. 2002. Immunophenotypic clustering of myelodysplastic syndromes. *Blood* 100, p. 2349.

McKusick, V. A. and Ruddle, F. H. 1987. Toward a complete map of the human genome. *Genomics* 1(2), pp. 103-106.

Megyeri, A. et al. 2005. Development of a stereological method to measure levels of fluoropyrimidine metabolizing enzymes in tumor sections using laser scanning cytometry. *Cytometry A* 64(2), pp. 62-71.

Meikrantz, W. and Schlegel, R. 1995. Apoptosis and the cell cycle. *J Cell Biochem* 58(2), pp. 160-174.

Melnikova, I. 2005. Future of COX2 inhibitors. *Nat Rev. Drug Discov.* 4, pp. 453-457.

Mendes, P. 1993. GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Comput Appl Biosci* 9(5), pp. 563-571.

Mendes, P. and Kell, D. 1997. Making cells work--metabolic engineering for everyone. *Trends Biotechnol* 15(1), pp. 6-7.

Minguez, J. M. et al. 2002. Synthesis and high content cell-based profiling of simplified analogues of the microtubule stabilizer (+)-discodermolide. *Mol Cancer Ther* 1(14), pp. 1305-1313.

Mittag, A. et al. 2005a. Polychromatic (eight-color) slide-based cytometry for the phenotyping of leukocyte, NK, and NKT subsets. *Cytometry A* 65(2), pp. 103-115.

Mittag, A. et al. 2005b. Polychromatic (eight-color) slide-based cytometry for the phenotyping of leukocytes, NK andf NKT subsets. *Cytometry Part A* 65(A), pp. 103-104.

Montalenti, F. et al. 1998. Simulating cancer-cell kinetics after drug treatment: Application to cisplatin on ovarian carcinoma. *PHYSICAL REVIEW E* 57(5), pp. 5877-5887.

Moraru, II et al. 2002. The virtual cell: an integrated modeling environment for experimental and computational cell biology. *Ann N Y Acad Sci* 971, pp. 595-596.

Murphy, R. et al. eds. 2005. IEEE transaction on image processing

Murray, A. and Hunt, T. 1993. *The Cell Cycle*. New York: Oxford Univ. Press, p. 251.

Nadkarni, P. M. 2002. An introduction to information retrieval: applications in genomics. *Pharmacogenomics J* 2(2), pp. 96-102.

Nelson, D. M. et al. 2002a. Coupling of DNA synthesis and histone synthesis in S phase independent of cyclin/cdk2 activity. *Mol Cell Biol* 22(21), pp. 7459-7472.

Nelson, G. et al. 2002b. Multi-parameter analysis of the kinetics of NF-kappaB signalling and transcription in single living cells. *J Cell Sci* 115(Pt 6), pp. 1137-1148.

Nelson, G. et al. 2002c. Dynamic analysis of STAT6 signalling in living cells. *FEBS Lett* 532(1-2), pp. 188-192.

Nigg, E. A. 1995. Cyclin-dependent protein kinases: key regulators of the eukaryotic cell cycle. *Bioessays* 17(6), pp. 471-480.

Noble, D. 2002a. Modeling the Heart--from Genes to Cells to the Whole Organ. *Science* 295(5560), pp. 1678-1682.

Noble, D. 2002b. The rise of computational biology. *Nat Rev Mol Cell Biol* 3(6), pp. 459-463.

Noctor, S. C. et al. 2001. Neurons derived from radial glial cells establish radial units in neocortex. *Nature* 409(6821), pp. 714-720.

Norbury, C. and Nurse, P. 1992. Animal cell cycles and their control. *Annu Rev Biochem* 61, pp. 441-470.

Novak, B. et al. 1999. Finishing the cell cycle. *J Theor Biol* 199(2), pp. 223-233.

Nurse, P. 2000a. The incredible life and times of biological cells. *Science* 289(5485), pp. 1711-1716.

Nurse, P. 2000b. A long twentieth century of the cell cycle and beyond. *Cell* 100(1), pp. 71-78.

O'Mahonya, R. et al. 2005. Comparison of Image Analysis software packages in the assessment of adhesion of micro-organisms to mucosal epithelium using confocal laser scanning microscopy. *J Micro meth* 61, pp. 105-126.

P´erez-Vel´azquez, J. et al. eds. 2008. *17th IFAC World Congress* Seoul, Korea:

Palaniappan, K. et al. eds. 2004a. *CVPR -IEEE Workshop on Articulated and Nonrigid Motion* Washington, DC,:

Palaniappan, K. et al. eds. 2004b. *IEEE Computer Vision and Pattern Recognition Workshop on Articulated and Nonrigid Motion* Washington, DC: IEEE Computer Society Press

Palkova, Z. et al. 2004. Single-cell analysis of yeast, mammalian cells, and fungal spores with a microfluidic pressure-driven chip-based system. *Cytometry Part A* 59(A), p. 246.

Palmer, E. and Freeman, T. 2005. Cell-based microarrays: current progress, future prospects. *Pharmacogenomics* 6(5), pp. 527-534.

Pap, E. H. et al. 1999. Ratio-fluorescence microscopy of lipid oxidation in living cells using C11-BODIPY(581/591). *FEBS Lett* 453(3), pp. 278-282.

Parker, L. L. and Piwnica-Worms, H. 1992. Inactivation of the p34cdc2-cyclin B complex by the human WEE1 tyrosine kinase. *Science* 257(5078), pp. 1955-1957.

Pawley, J., ed. 1989. *The Handbook of Biological Confocal Microscopy*. Madison, WI: IMR Press, p. 201.

Pearson, W. R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol* 183, pp. 63-98.

Pearson, W. R. and Lipman, D. J. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* 85(8), pp. 2444-2448.

Perfetto, S. P. et al. 2004. Seventeen-colour flow cytometry: unravelling the immune system. *Nat Rev Immunol* 4(8), pp. 648-655.

Perlman, Z. E. et al. 2004a. Multidimensional drug profiling by automated microscopy. *Science* 306(5699), pp. 1194-1198.

Perlman, Z. E. et al. 2004b. Multidimensional drug profiling by automated microscopy. *Science* 306, pp. 1194-1196.

Pierrez, J. and Ronot, X. 2004. Flow cytometric analysis of the cell cycle: Mathematical modeling and biological interpretation. *Acta Biotheoretica* 40, pp. 131-137.

Pines, J. and Hunter, T. 1994. The differential localization of human cyclins A and B is due to a cytoplasmic retention signal in cyclin B. *Embo J* 13(16), pp. 3772-3781.

Pollok, B. 2005. Assay development: an increasingly creative endeavour. *Nature Reviews Drug Discovery* 4, pp. 956-957.

Pomerening, J. R. et al. 2005. Systems-level dissection of the cell-cycle oscillator: bypassing positive feedback produces damped oscillations. *Cell* 122(4), pp. 565-578.

Pomerening, J. R. et al. 2003. Building a cell cycle oscillator: hysteresis and bistability in the activation of Cdc2. *Nat Cell Biol* 5(4), pp. 346-351.

Pommier, Y. 2006. Topoisomerase I inhibitors: camptothecins and beyond. *Nat Rev Cancer* 6(10), pp. 789-802.

Pommier, Y. et al. 2004. Apoptosis defects and chemotherapy resistance: molecular interaction maps and networks. *Oncogene* 23(16), pp. 2934-2949.

Potel, M. J. et al. 1979. A system for interactive film analysis. *Comput Biol Med* 9(3), pp. 237-256.

Price, J. H. et al. 2002. Advances in molecular labeling, high throughput imaging and machine intelligence portend powerful functional cellular biochemistry tools. *J Cell Biochem Suppl* 39, pp. 194-210.

Prieur-Carrillo, G. et al. 2003. Computerized Video Time-Lapse (CVTL) Analysis of the Fate of Giant Cells Produced by X-Irradiating EJ30 Human Bladder Carcinoma Cells. *Rad. Res.* 159, pp. 705-712.

Psaty, B. M. et al. 2004. Potential for conflict of interest in the evaluation of suspected adverse drug reactions: Use of cerivastatin and risk of rhabdomyolysis. *JAMA* 292, pp. 2622-2634.

Qu, Z. et al. 2003. Regulation of the mammalian cell cycle: a model of the G1-to-S transition. *Am J Physiol Cell Physiol* 284(2), pp. C349-364.

Raff, J. W. et al. 2002. The roles of Fzy/Cdc20 and Fzr/Cdh1 in regulating the destruction of cyclin B in space and time. *J Cell Biol* 157(7), pp. 1139-1149.

Rainsford, K. D. 2007. Anti-inflammatory drugs in the 21st century. *Subcell Biochem* 42, pp. 3-27.

Rashid, S. T. et al. 2004. Engineering of bypass conduits to improve patency. *Cell Prolif* 37(5), pp. 351-366.

Reif, D. M. et al. 2004. Integrated analysis of genetic, genomic and proteomic data. *Expert Rev Proteomics* 1(1), pp. 67-75.

Reits, E. A. et al. 1997. Dynamics of proteasome distribution in living cells. *Embo J* 16(20), pp. 6087-6094.

Rieder, C. L. and Khodjakov, A. 2003. Mitosis through the microscope: advances in seeing inside live dividing cells. *Science* 300, pp. 91-96.

Rines, D. et al. 2006. High-content screening of functional genomic libraries. *Methods Enzymol* 414, pp. 530-565.

Ronot, X. et al. 2000. Quantitative study of dynamic behavior of cell monolayers during in vitro wound healing by optical flow analysis. *Cytometry* 41(1), pp. 19-30.

Roques, E. J. and Murphy, R. F. 2002. Objective evaluation of differences in protein subcellular distribution. *Traffic* 3(1), pp. 61-65.

Rosa, J. et al. 2006. Survivin modulates microtubule dynamics and nucleation throughout the cell cycle. *Mol Biol Cell* 17(3), pp. 1483-1493.

Sahai, E. et al. 2005. Simultaneous imaging of GFP, CFP and collagen in tumors in vivo using multiphoton microscopy. *BMC Biotechnol* 5, p. 14.

Salmon, E. D. 1995. VE-DIC light microscopy and the discovery of kinesin. *Trends Cell Biol* 5(4), pp. 154-158.

Sampath, D. and Plunkett, W. 2001. Design of new anticancer therapies targeting cell cycle checkpoint pathways. *Curr Opin Oncol* 13(6), pp. 484-490.

Sams-Dodd, F. 2005. Target-based drug discovery: is something wrong? . *Drug Dev. Tech.* 10, p. 139.

Sauro, H. et al. 2003. Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. *OMICS* 7(4), pp. 355-372.

Sawai, S. et al. 2007. High-throughput analysis of spatio-temporal dynamics in Dictyostelium. *Genome Biol* 8(7), p. R144.

Schadt, E. E. and Lum, P. Y. 2006. Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Reverse engineering gene networks to identify key drivers of complex disease phenotypes. *J Lipid Res* 47(12), pp. 2601-2613.

Schleiden, M. 1838. *Arch. Anat. Physiol* 13, p. 137–176.

Schneider, M. 2004. A rational approach to maximize success rate in target discovery. *Arch Pharm* 337, p. 625.

Schubert, W. 1990. Multiple antigen-mapping microscopy of human tissue. In: Burger, G. et al. eds. *Advances in Analytical Cellular Pathology.* Amsterdam: Excerpta Medica, p. 97.

Schubert, W. 2004. Exploring the toponome: localize hundred proteins (or more) at once and walk through protein networks, cell by cell. *Cytometry Part A* 64(A), pp. 117-119.

Schwann, T. 1839. *Mikroskopische Untersuchungen über die Übereinstimmung in der Struktur und dem Wachstum der Tiere und Pflanzen* Sander'schen Buchhandlung.

Schwartzman, R. A. and Cidlowski, J. A. 1993. Apoptosis: the biochemistry and molecular biology of programmed cell death. *Endocr Rev* 14(2), pp. 133-151.

Shaner, N. C. et al. 2008. Improving the photostability of bright monomeric orange and red fluorescent proteins. *Nat Methods* 5(6), pp. 545-551.

Shapiro, L. and Losick, R. 1997. Protein localization and cell fate in bacteria. *Science* 276(5313), pp. 712-718.

Shen, C. et al. 2008. Patterning Cell and Tissue Function. *Cellular and Molecular Bioengineering* 1, pp. 15-23.

Shen, F. and Price, J. H. 2006. Toward complete laser ablation of melanoma contaminant cells in a co-culture outgrowth model via image cytometry. *Cytometry A* 69(7), pp. 573-581.

Shen, H. et al. 2006. Automatic tracking of biological cells and compartments using particle filters and active contours. *Chem Intel Laboratory Sys* 82, pp. 276 - 282.

Sible, J. C. and Tyson, J. J. 2007. Mathematical modeling as a tool for investigating cell cycle control networks. *Methods* 41(2), pp. 238-247.

Simpson, A. J. 2001. Genome sequencing networks. *Nat Rev Genet* 2(12), pp. 979-983.

Slater, K. 2001. Cytotoxicity tests for high-throughput drug discovery. *Curr Opin Biotechnol* 12(1), pp. 70-74.

Smith, J. A. and Martin, L. 1973. Do Cells Cycle? *Proceedings of the National Academy of Sciences* 70(4), pp. 1263-1267.

Smith, P. et al. 2008. Cell cycle checkpoint-guarded routes to catenation-induced chromosomal instability. *SEB Exp Biol Ser.* 59, pp. 219-242.

Smith, P. J. 2004. Cytomics—Drug discovery via cytometry in time and space. Business Briefing PharmaTech 2004

Smith, P. J. et al. 2000. Characteristics of a novel deep red/infrared fluorescent cell-permeant DNA probe, DRAQ5, in intact human cells analyzed by flow cytometry, confocal and multiphoton microscopy. *Cytometry* 40(4), pp. 280-291.

Smith, P. J. et al. 2007a. Cytomics and drug development. *Cytometry A* 71(6), pp. 349-351.

Smith, P. J. et al. 2007b. Mitotic bypass via an occult cell cycle phase following DNA topoisomerase II inhibition in p53 functional human tumor cells. *Cell Cycle* 6(16), pp. 2071-2081.

Sommer, R. J. et al. 1994. The evolution of cell lineage in nematodes. *Dev Suppl*, pp. 85-95.

Spearman, C. 1904. The proof and measurement of association between two things. *Amer. J. Psychol* 15, pp. 72-101.

Steel, G. 1977. Growth kinetics of tumors.*Cell Population Kinetics in Relation to the Growth and Treatment of Cancer.* Oxford: Clarendon Press.

Stefansson, B. and Brautigan, D. L. 2007. Protein phosphatase PP6 N terminal domain restricts G1 to S phase progression in human cancer cells. *Cell Cycle* 6(11), pp. 1386-1392.

Steiner, G. E. et al. 2000. Automated data acquisition by confocal laser scanning microscopy and image analysis of triple stained immunofluorescent leukocytes in tissue. *J Immunol Methods* 237(1-2), pp. 39-50.

Stephens, P. et al. 2004. Crosslinking and G-protein functions of transglutaminase 2 contribute differentially to fibroblast wound healing responses. *J Cell Sci* 117(Pt 15), pp. 3389-3403.

Stern, C. D. and Fraser, S. E. 2001. Tracing the lineage of tracing cell lineages. *Nat Cell Biol* 3(9), pp. E216-218.

Sternberg, P. W. and Horvitz, H. R. 1981. Gonadal cell lineages of the nematode Panagrellus redivivus and implications for evolution by the modification of cell lineage. *Dev Biol* 88(1), pp. 147-166.

Sternberg, P. W. and Horvitz, H. R. 1982. Postembryonic nongonadal cell lineages of the nematode Panagrellus redivivus: description and comparison with those of Caenorhabditis elegans. *Dev Biol* 93(1), pp. 181-205.

Stirland, J. A. et al. 2003. Real-time imaging of gene promoter activity using an adenoviral reporter construct demonstrates transcriptional dynamics in normal anterior pituitary cells. *J Endocrinol* 178(1), pp. 61-69.

Stransky, B. et al. 2007. Modeling cancer: integration of "omics" information in dynamic systems. *J Bioinform Comput Biol* 5(4), pp. 977-986.

Sturn, A. et al. 2002. Genesis: cluster analysis of microarray data. *Bioinformatics* 18(1), pp. 207-208.

Stywester, D. and Dennis, S. 1980. COMPUTER PROCESSING OF CELL LINEAGE DATA FROM TIME LAPSE CINEMATOGRAPHY STUDIES. *Compur. Biol. Med* 10, pp. 103-108.

Swedlow, J. R. et al. 2003. Informatics and quantitative analysis in biological imaging. *Science* 300(5616), pp. 100-102.

Systems Biology Report 2007. Systems Biology: a vision for engineering and medicine. In: sciences, T.R.A.o.E.a.T.R.A.o.M. ed. London:

Szaniszlo, P. et al. 2004. Getting the right cells to the array: Gene expression microarray analysis of cell mixtures and sorted cells. *Cytometry A* 59(2), pp. 191-202.

Tang, M. et al. 2003. Microsatellite analysis of synchronous and metachronous tumors: a tool for double primary tumor and metastasis assessment. *Diagn Mol Pathol* 12(3), pp. 151-159.

Tatebe, H. et al. 2001. Fission yeast living mitosis visualized by GFP-tagged gene products. *Micron* 32(1), pp. 67-74.

Taylor, C. F. et al. 2008. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol* 26(8), pp. 889-896.

Taylor, D. 2006. Past, Present, and Future of High Content Screening and the Field of Cellomics. In: Taylor, D. et al. eds. *Methods in Molecular Biology.* Vol. 356. New Jersey: Humana Press, Inc, pp. 3-18.

Taylor, D. et al. 1984. Fluorescent analog cytochemistry. *Trends in Biochemical Science* 9, pp. 88-91.

Taylor, D. and Giuliano, K. 2005. Multiplexed high content screening assays create a systems cell biology approach to drug discovery. *Drug Discov Today Technologies* 2(2), pp. 149-154.

Taylor, D. L. and Salmon, E. D. 1989. Basic fluorescence microscopy.*Methods in Cell Biology.* Vol. 29. New York: Academic, pp. 208-237.

Taylor, D. L. et al. 1986. *Applications of Fluorescence in the Biomedical Sciences.* New York: Liss, p. 639.

Taylor, D. L. et al. 2001. Real-time molecular and cellular analysis: the new frontier of drug discovery. *Curr Opin Biotechnol* 12(1), pp. 75-81.

Taylor, T. B. et al. 2004. Microgenomics: identification of new expression profiles via small and single-cell sample analysis. *Cytometry Part A* 59(A), pp. 254-257.

Terstappen, G. et al. 2007. Target deconvolution strategies in drug discovery *Nature Reviews Drug Discovery* 6, pp. 891 - 903.

Thearling, K. 2008. An Introduction to Data Mining.

Thomas, N. 2003. Lighting the circle of life: fluorescent sensors for covert surveillance of the cell cycle. *Cell Cycle* 2(6), pp. 545-549.

Thomas, N. and Goodyear, I. 2003. Stealth sensors: real-time monitoring of the cell cycle. *Targets* 2, pp. 26-33.

Tsien, R. Y. 1989. Fluorescent probes of cell signaling. *Annual Review of Neuroscience* 12, pp. 227-253.

Tsien, R. Y. 1998. The green fluorescent protein. *Annu Rev Biochem* 67, pp. 509-544.

Tsien, R. Y. and Waggoner, A. S. 1990. Fluorophores for confocal microscopy: photophysics and photochemistry. In: Pawley, J. ed. *he Handbook of Biological Confocal Microscopy.* Madison, WI: IMR Press.

Twyman, R. 2004. *Principles of proteomics.* New York: BIOS Scientific Publishers.

Tyson, J. 1974/75. On the existence of oscillatory solutions in negative feedback cellular control processes. *J. Math. Biol* 1, pp. 311-315.

Tyson, J. 2002. Cell cycle controls. In: Fall, C. et al. eds. *Computational Cell Biology.* New York, Berlin: Springer, pp. 261-284.

Tyson, J. and Sachsenmaier, W. 1978. Is nuclear division in Physarum controlled by a continuous limit cycle oscillator? *J. Theor. Biol.* 73, pp. 723-738.

Tyson, J. J. 1991. Modeling the cell division cycle: cdc2 and cyclin interactions. *Proc Natl Acad Sci U S A* 88(16), pp. 7328-7332.

Tyson, J. J. and Novak, B. 2001. Regulation of the eukaryotic cell cycle: molecular antagonism, hysteresis, and irreversible transitions. *J Theor Biol* 210(2), pp. 249-263.

Ubezio, P. and Rossotti, A. 1987. Sensitivity of flow cytometric data to variations in cell cycle parameters. *Cell Prolif.* 20(5), pp. 507-517.

Uetz, P. et al. 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 403(6770), pp. 623-627.

Valet, G. 2005a. Cytomics, the human cytome project and systems biology: top-down resolution of the molecular biocomplexity of organisms by single cell analysis. *Cell Prolif.* 38, pp. 171-174.

Valet, G. 2005b. Cytomics: an entry to biomedical cell systems biology. *Cytometry A* 63(2), pp. 67-68.

Valet, G. et al. 2004. Cytomics—New Technologies: Towards a Human Cytome Project. *Cytometry Part A* 59(A), pp. 167-171.

Valet, G. et al. 2003. Pretherapeutic identification of high-risk acute myeloid leukemia (AML) patients from immunophenotypic, cytogenetic, and clinical parameters. *Cytometry Part B Clin Cytom* 53(B), p. 4.

Van den Heuvel, S. and Harlow, E. 1993. Distinct roles for cyclin-dependent kinases in cell cycle control. *Science* 262(5142), pp. 2050-2054.

Van Osta, P. 2006. Extracting quantitative information from tissue--an industrial perspective. *Cytometry A* 69(7), pp. 588-591.

Van Osta, P. et al. 2006. Cytomics and drug discovery. *Cytometry A* 69(3), pp. 117-118.

Vancoppenolle, B. et al. 2000. Evaluation of fixation methods for ultrastructural study of Caenorhabditis elegans embryos. *Microsc Res Tech* 49(2), pp. 212-216.

Walker, M. G. 2001. Drug target discovery by gene expression analysis: cell cycle genes. *Curr Cancer Drug Targets* 1(1), pp. 73-83.

Wang, Y.-L. and Taylor, D. L. 1989. Fluorescence Microscopy of Living Cells in Culture, Part A: Fluorescent Analogs, Labeling Cells and Basic Microscopy.*Methods in Cell Biology.* Vol. 29. New York: Academic, p. 333.

Watson, J. 1991. *Introduction to Flow Cytometry* Cambridge: Cambridge University Press.

Weigelt, B. et al. 2003. Gene expression profiles of primary breast tumors maintained in distant metastases. *Proc Natl Acad Sci U S A* 100(26), pp. 15901-15905.

Weingartner, M. et al. 2001. Dynamic recruitment of Cdc2 to specific microtubule structures during mitosis. *Plant Cell* 13(8), pp. 1929-1943.

Weston, A. D. and Hood, L. 2004. Systems biology, proteomics, and the future of health care: towards predictive, preventatitve, and personalized medicine. . *J. Proteome Res.* 3, p. 179.

White, M. et al. 2005. Real-time imaging of cell division and apoptosis. *Toxicology* 213(3), pp. 205-206.

White, N. S. and Errington, R. J. 2005. Fluorescence techniques for drug delivery research: theory and practice. *Adv Drug Deliv Rev* 57(1), pp. 17-42.

Whitman, C. 1878. The embryology of Clepsine. *Q. J. Morphol. (N.S.)* 18, pp. 215-315.

Whitman, C. 1887. A contribution to the history of germ layers in Clepsine. *J. Morphol* 1, pp. 105-182.

Wiegner, O. and Schierenberg, E. 1998. Specification of gut cell fate differs significantly between the nematodes Acrobeloides nanus and caenorhabditis elegans. *Dev Biol* 204(1), pp. 3-14.

Wipf, P. et al. 2000. Synthesisand biological evaluation of a focused mixture library of analogues ofthe antimitotic marine natural product uracin. *A. J. Am. Chem. Soc* 122, pp. 9391-9395.

Wishart, D. S. 2007. Human Metabolome Database: completing the 'human parts list'. *Pharmacogenomics* 8(7), pp. 683-686.

Wodarz, A. and Gonzalez, C. 2006. Connecting cancer to the asymmetric division of stem cells. *Cell* 124(6), pp. 1121-1123.

Wood, W. 1999. Cell lineages in Caenorhabditis elegans development. In: Moody, S. ed. *Cell Lineage and Fate Determination.* San Diego: Academic Press, pp. 77-95.

Wu, H. et al. 2004. Chemical cytometry on a picoliter-scale integrated microfluidic chip. *PNAS* 101, p. 12809.

Wu, R. S. and Bonner, W. M. 1981. Separation of basal histone synthesis from S-phase histone synthesis in dividing cells. *Cell* 27(2 Pt 1), pp. 321-330.

Wu, X. et al. 2008. Network-based global inference of human disease genes. *Mol Syst Biol* 4, p. 189.

Yamamoto, N. et al. 2003. Determination of clonality of metastasis by cell-specific color-coded fluorescent-protein imaging. *Cancer Res* 63(22), pp. 7785-7790.