

INTERNATIONAL JOURNAL OF KNOWLEDGE DISCOVERY IN BIOINFORMATICS

October-December 2010, Vol. 1, No. 4

Table of Contents

IEEE ICDM WORKSHOP ON BIOLOGICAL DATA MINING AND ITS
APPLICATIONS IN HEALTHCARE

EDITORIAL PREFACE

- i *Xiao-Li Li, Institute for Infocomm Research, Singapore*
See-Kiong Ng, Institute for Infocomm Research, Singapore

RESEARCH ARTICLES

- 1 **Hierarchical Density-Based Clustering of White Matter Tracts in the Human Brain**
Junming Shao, University of Munich, Germany
Klaus Hahn, HMGU Helmholtz Center Munich, Germany
Qinli Yang, University of Edinburgh, UK
Afra Wohlschlaeger, Technical University of Munich, Germany
Christian Boehm, University of Munich, Germany
Nicholas Myers, Technical University of Munich, Germany
Claudia Plant, Florida State University, USA
- 26 **Revealing the Origin and Nature of Drug Resistance of Dynamic Tumour Systems**
Ricardo Santiago-Mozos, National University of Ireland Galway, Ireland
Imtiaz A. Khan, Cardiff University, UK
Michael G. Madden, National University of Ireland Galway, Ireland
- 54 **Improving Prediction Accuracy via Subspace Modeling in a Statistical Geometry Based Computational Protein Mutagenesis**
Majid Masso, George Mason University, USA
- 69 **Efficient Mining Frequent Closed Discriminative Biclusters by Sample-Growth: The FDCluster Approach**
Miao Wang, Northwestern Polytechnical University, China
Xuequn Shang, Northwestern Polytechnical University, China
Shaohua Zhang, Northwestern Polytechnical University, China
Zhanhuai Li, Northwestern Polytechnical University, China

Revealing the Origin and Nature of Drug Resistance of Dynamic Tumour Systems

Ricardo Santiago-Mozos, National University of Ireland Galway, Ireland

Imtiaz A. Khan, Cardiff University, UK

Michael G. Madden, National University of Ireland Galway, Ireland

ABSTRACT

In this paper, the authors identify the strategies that resistant subpopulations of cancer cells undertake to overcome the effect of the anticancer drug Topotecan. For the analyses of cell lineage data encoded from timelapse microscopy, data mining tools are chosen that generate interpretable models of the data, addressing their statistical significance. By interpreting the short-term and long-term cytotoxic effect of Topotecan through these data models, the authors reveal the strategies that resistant subpopulations of cells undertake to maximize their clonal expansion potential. In this context, this paper identifies a pattern of cell death independent of cytotoxic effect. Finally, it is observed that cells exposed to Topotecan have higher movement over time, indicating a putative relationship between cytotoxic effect and cell motility.

Keywords: Cancer, Cell Lineage, Cytomics, Data Mining, Drug Resistance, Oncology, Osteosarcoma, Topotecan

A central challenge in cancer biology is to understand of the strategies followed by cells to overcome the effect of anticancer drugs (Smith, Khan, & Errington, 2009). Cancer, a highly complex and heterogeneous disease (Heppner, 1984; Rubin, 1990), can be described as an evolving system, that can be best illustrated by cell lineages. In experimental terms, a cell lineage reflects the relationship between descendents from a common progenitor that was exposed to a given influence, such as a bioactive drug, for a time period. The behaviour of both the pro-

genitor and the evolving progeny reveals the time-integrated response to this influence (i.e., the pharmacodynamic response). The study of cell lineages has been, and remains, of importance in developmental biology (Stern & Fraser, 2001; Alvarez-Buylla, García-Verdugo, & Tramontin, 2001; Anderson, Gage, & Weissman, 2001; Ardavin et al., 2001; Dor, Brown, Martinez, & Melton, 2004; Kim & Shibata, 2002; Noctor, Flint, Weissman, Dammerman, & Kriegstein, 2001) and medicine (Bernards & Weinberg, 2002; Hope, Jin, & Dick, 2004; Tang et al., 2003; Weigelt et al., 2003; Yamamoto et al., 2003).

DOI: 10.4018/jkdb.2010100102

Here we select and apply appropriate data mining techniques that provide interpretable models on previously encoded cell lineage data (Khan et al., 2007), in order to reveal the degree of heterogeneity of a tumour system in response to therapy, as well as the strategies (or patterns) the resistant fraction incorporates in order to overcome the effect of the anticancer drug and thus maximize their clonal expansion potential.

DATA

Biological Sample Preparation

The human osteosarcoma cell line [U-2 OS (ATCC HTB 96)23], derived from a 15-year-old Caucasian female and transfected with a fluorescent reporter cyclin B1 GFP, is selected. The cells are maintained at 37 °C and 5% CO₂ using standard tissue culture techniques. Media used is McCoy's 5A modified (Sigma) supplemented with 2 mM glutamine, 100 units/ml penicillin, 100 mg/ml streptomycin, 10% fetal calf serum and 1000 mg/ml geneticin.

Cells are treated with 1 μM and 10 μM bolus dose of anticancer agent Topotecan (TPT) (Bailey, 2000). TPT is a water soluble derivative of the alkaloid Camptothecin and act as a topoisomerase I, a nuclear enzyme involved in DNA replication and repair, inhibitor (Wang, 1996). TPT is used for the treatment of a wide range of cancers, including lung (Ichinose et al., 2010), breast (Cheung et al., 2008), ovarian (Lorusso et al., 2010) and bone (Seibel et al., 2007), both in experimental and clinical contexts.

An hour post treatment, the cultured dishes are placed onto a time lapse instrument designed to capture transmission phase images from multi well plates. Image sequences are taken for 115 hours at 15-minute time intervals. The cell lineage data is encoded by ProgeniTRAK (Khan et al., 2006) and is retrieved from a cell lineage database ProgeniDB (Khan et al., 2007).

Cell Lineage Data

Information from 253 lineages is available: 168 are Control, 37 are 1 μM and 48 are 10 μM

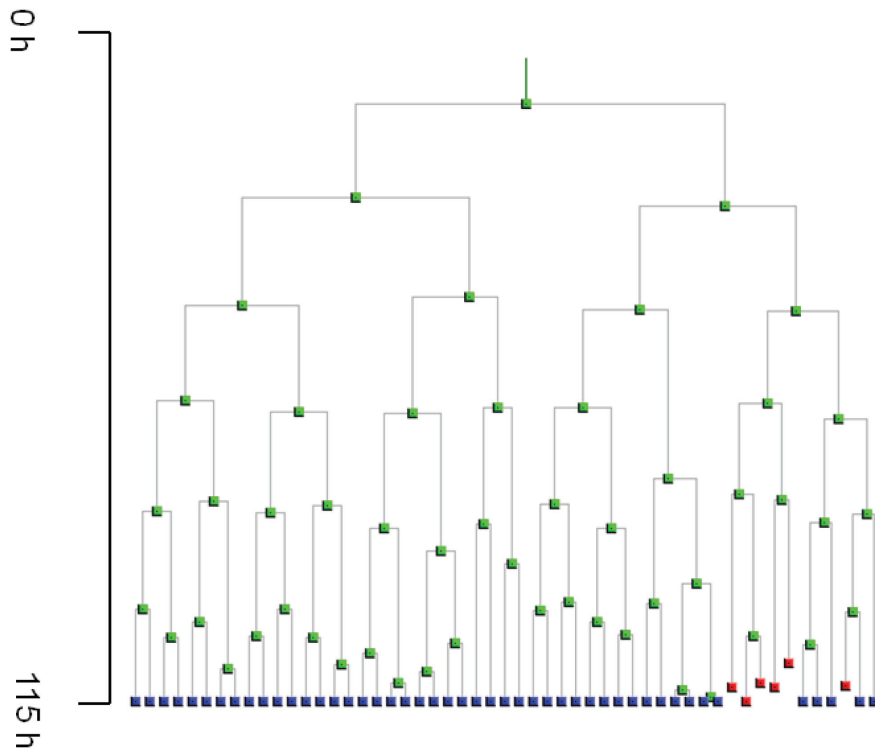
lineages. Each lineage is characterised as a tree where the genealogy of the Progenitor cell and its offspring is represented.

Figure 1 shows an example of a Control lineage along the experiment; we observe that Sixth Generation is reached in the 115 hours. Figure 2 shows the encoding, along with the labeling convention (refer to Table 1 for notation): an arc between two successive nodes, where both nodes represents dividing (mitosis) cells, indicates the cell cycle time or inter-mitotic-time (IMT). In labeling terms, for example, Cell_{3,1} and Cell_{3,2} in Figure 2 are sisters; both are daughters of Cell_{2,1}, which is a daughter of Cell_{1,1}, which is again a daughter of a Progenitor cell. The daughters of a Progenitor are termed First Generation cells (Cell_{1,1} and Cell_{1,2} in Figure 2). The daughters of the First Generation cells are termed Second Generation cells (Cell_{2,1}, ..., Cell_{2,4} in Figure 2), and so on. For this work, we consider nodes or cells up to the Fourth Generation, with the assumption that any living cells after Third Generation can be considered to be part of the resistant population. By the Third Generation, due to the amount of time elapsed since the administration of the drug and the cell proliferation rate, the cytotoxic effect of the drug is diminished.

We record the status of each cell in the lineage; see Table 2 for possible statuses and their definitions. When a cell divides, both the IMT and its inter-mitotic displacement (IMD) are recorded. IMT refers to the time taken by a cell to divide, while IMD refers to its displacement during that time. The time needed by the Progenitor cell to divide is recorded as DIV and its displacement from the start of the experiment until it divides is recorded as DIV_D.

Data Transformation. Sister cells are sorted according their statuses (following the order given in Table 2) and when both cells divide, the one with shorter IMT goes first. For example, if both First Generation cells divide, the one with shorter IMT is labeled as Cell_{1,1} and the other as Cell_{1,2}. The daughters of Cell_{1,1} are called Cell_{2,1} and Cell_{2,2}, and are sorted according their

Figure 1. A lineage encoded from a real progenitor cell using ProgeniTRAK, spanning up to six generations, where each green node represents cell division, red node cell death



statuses and IMTs as before. IMT and IMD numbering follow their cell numbering; for example, $IMT_{3,5}$ refers to the IMT of $Cell_{3,5}$ (see Table 1). Lineage information is codified in a vector of 44 variables:

1. DIV, DIV_D : all progenitor cells divide in our dataset.
2. $Status_{1,j}$ (14 variables, one for each cell apart from the Progenitor cell; see Table 2 for their possible values).
3. $IMT_{1,j}$ and $IMD_{1,j}$ (28 variables): these are numeric values except if a cell does not divide, in which case the *missing* value is recorded.

Implications. The sorting of the data have some implications: since *Divided* is the last status in the ordering, if the first sister divides, the second sister also divides, and has a larger

IMT. Of course, if the second sister does not divide, neither does the first and their statuses are dependent according the ordering in Table 2. For example, If $Status_{3,5}$ equals *Divided* that implies that $Cell_{3,6}$ also divided; and it implies that $Cell_{2,3}$ divided (its mother) and this last fact implies that $Cell_{2,4}$ divided; and finally than $Cell_{1,2}$ divided. Therefore, when we say that $Cell_{3,5}$ divided we are implicitly saying that $Cell_{1,2}, Cell_{2,3}, Cell_{2,4}$ and $Cell_{3,6}$ also divided, which makes this a very informative event.

STUDIES

The data are analysed to understand the behaviour of Control lineages, and we explore any deviations from this behaviour that arise when the drug is applied. We have selected four main studies, as described below: (1) cell death in Control lineages; (2) cytotoxic effect

Figure 2. Generalized labeling convention for transforming real lineages into machine learning terms. For this work we included lineage data up to the third generation.

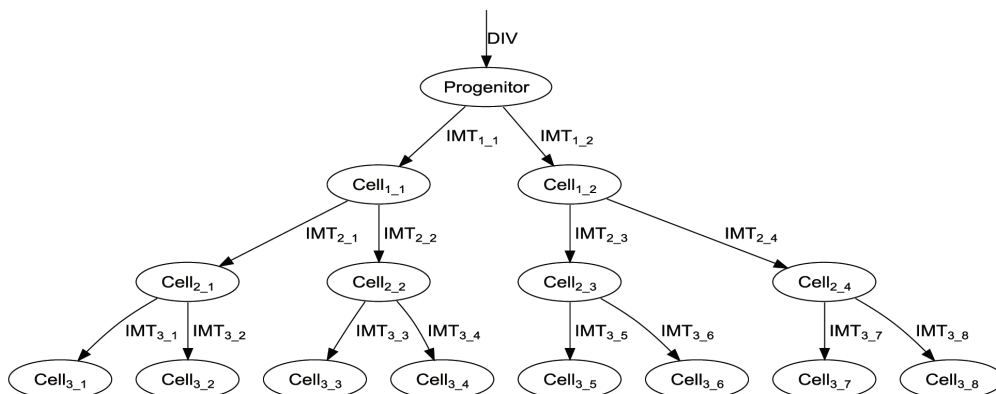


Table 1. Notation and definitions

Variable	Definition
Cell _{I,J}	Cell J in Generation I.
IMT	Inter-Mitotic Time. The time needed by a cell to divide.
IMD	Inter-Mitotic Displacement. The distance travelled by a cell until its division.
DIV	The elapsed time from the start of the experiment until division of a Progenitor cell.
DIV_D	The distance travelled by a Progenitor cell from the start of the experiment until division.
Status _{I,J}	Status of Cell J in Generation I.
IMT _{I,J}	The IMT for cell J in Generation I.
IMD _{I,J}	The IMD for cell J in Generation I.
Resistant	Cells whose descendants reach Fourth Generation. When applied to lineages, lineages that reach Fourth Generation.
Stressed	The opposite to resistant.
Control	Cells (lineages) that have not been exposed to drug.
Perturbed	Cells (lineages) that have been exposed to drug.
Drug	A variable that carries information to distinguish the type of lineage.

of the drug; (3) characterization of Stressed lineages; and (4) classification of Control versus Perturbed lineages.

Study of Cell Death in Control Lineages

The goal of this study is to explore *Death* events in Control lineages. The question considered is:

‘is there any discernible pattern in the *Death events*?’ For this analysis, we divide the cells into two groups: the first contains the mothers of cell sisters where at least one sister dies, while in the second contains the mothers of cell sisters where both of the sisters survive. For each generation, the cell *death factor*, defined as the ratio of *Death* cells to *Divided* cells,

Table 2. Possible statuses for cells

Cell status	Definition
M4	The cell divides into four daughters.
M3	The cell divides into three daughters.
Unresolved	The cell status could not be measured.
Polyploidy	The cell divides into one cell.
Re-fused	The cell divides but its daughters cells fuse together.
Arrested	The cell survives without dividing until the end of the experiment.
Lost	The cell is lost from the field of view of the microscope.
Never-Born	The mother of the cell does not divide.
Death	The cell dies due to apoptosis (programmed cell death) or necrosis (uncontrolled cell death).
Divided	Normal division. The cell divides into two daughters.

and the dynamical behaviour of the mothers of *Death* and *Non-Death* cells (those with any status except *Death*, *Never-Born* or *Lost*) are measured and compared. Also, we construct classifiers to predict whether a cell pair is going to have at least one *Death*, given their ancestors' information. Note that it is not possible in this classification task to distinguish between sisters, as they share the same ancestors' information, even though they can be different: for example, one might divide while the other dies. Accordingly, the task is inherently noisy.

Study of the Cytotoxic Effect of Topotecan

The goal of this study is to determine the effectiveness of the drug. We consider the difference of ratios of cell events in Control, 1 μ M and 10 μ M lineages, by focusing on the following measures:

1. The ratio of *Death Cells* to *Non-Death Cells* measures the catastrophic cytotoxic effect of the drug.
2. The ratio of *Stressed Cells* to *Resistant Cells*. Stressed cells are those that manage to divide but whose offspring does not have Third Generation dividing cells. Resistant

cells have an offspring with at least one Third Generation dividing cell. This ratio measures the cytotoxic effect on the current generation based on the outcome in the Third Generation, which can be interpreted as a moderated cytotoxic effect.

3. The ratio of *Full Clonal Potential Cells* (FCP) to *Partial Clonal Potential Cells* (PCP). FCP cells are those whose offspring in the Third Generation are all dividing cells. For cells at the First Generation, full clonal potential means that they have four descendants that divide in Third Generation, and obviously for cells in the Second Generation, full clonal potential means they have two descendants that divide in the Third Generation. Partial clonal potential cells are those whose offspring have *at least one* Third Generation dividing cell but are *not* full clonal potential. This proportion measures the cytotoxic effect on the current generation based on the outcome in Third Generation resistant cells; in other words, it measures the cytotoxic effect on the cells' clonal potential.

In addition, the effect of the drug on the cell dynamics (IMT and IMD) is investigated.

Study of Stressed Lineages

The goal of this study is to identify the common strategies followed by the cells to reach the Fourth Generation. Lineages are divided into a *Stressed* group, which do not reach this generation, and a *Resistant* group (the others). Then, classification rules that identify both groups are induced.

Study of Control versus Perturbed Lineages

When the drug is applied, cell lineages are expected to be perturbed. The goal of this study is to identify what lineages have been perturbed. That is, what differences can be identified between Control and Perturbed lineages. For this study, the data is separated into two classes, one containing the Control lineages and the other containing both 1 μ M and 10 μ M lineages.

METHODS

To compare proportions of events among lineage types, the events are counted and arranged in contingency tables, and the Fisher test is used to assess if the difference between the proportions is statistically significant (Campbell, 2007). The null hypothesis is: *both lineage types have the same proportion on these events*, and the alternative hypothesis is: *the proportions are different*. As several tests are performed, the false discovery rate (FDR) must be controlled (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001). The FDR is the expected number of false rejections among all rejections of the null hypothesis. To control the FDR with Fisher tests, we use the method proposed in Carlson, Heckerman, and Shani (2009), with an FDR value of 0.1.

To compare the dynamic behaviours of lineages we select the empirical cumulative distribution function (ECDF), which approximates the probability that a variable is equal to or lower than any threshold. The ECDF converges to the actual cumulative distribu-

tion function as the number of samples tends to infinity (Vapnik, 1998). To compare ECDF distributions, the Kolmogorov-Smirnov two-sample test (Smirnov, 1939) is used. The null hypothesis is: *both samples come from a common distribution*, and we choose as the alternative hypothesis: *both samples come from different distributions*. To control the FDR of these tests, the general method proposed in (Benjamini & Yekutieli, 2001, Theorem 1.3) is used. The reason for using two different methods to control the FDR (one for the Fisher tests and other for the Kolmogorov-Smirnov test) is because the method proposed in Carlson et al. (2009) is more appropriate (and has more power) for contingency tables than the general procedure. To learn patterns from the data, we use RIPPER (Cohen, 1995), which is a rule inducer, and C4.5 (Quinlan, 1996), which is an algorithm to build tree classifiers, because these algorithms can provide interpretable representations of their classification decisions.

Data is also explored using Bayesian networks (Charniak, 1991; Pearl, 1988; Howard & Matheson, 2005). Continuous models are avoided because the distribution of the data is unknown. Instead, we consider the following methods for data discretization: equal-width bins, equal-frequency bins and the supervised discretization proposed in (Fayyad & Irani, 1993). After learning the Bayesian network from the data, the probability distributions of the variables given the values of their mothers provide valuable information about the data.

The performance of the classification algorithms is measured by leave-one-out cross-validation, and all processing steps in the data are done inside the leave-one-out loop. Finally, models that are presented for discussion in this paper are induced from all the data.

RESULTS

In this section, we show the results of applying the different methods to the data, with the aim of addressing the research questions identified above.

Study of Cell Death in Control Lineages

Table 3 summarises the results for cell death in Control lineages. There is a clear increase in the cell death factor in the Second Generation. However, the most important observation here is that cell death occurs spontaneously as part of the regulation of the population growth in the Control group. Given that there are 10 possible statuses and *Death* is not a particularly frequent one, an important pattern is found in death cells: *Death-Death* sisters pairs account for almost half of deaths in the First Generation (6 out of 13), more than half in the Second Generation (28 out of 51) and half in the Third Generation (20 out of 40).

The next analysis focuses on cell dynamics. IMT_1 (both IMT_{1_1} and IMT_{1_2}) is divided into two groups:

1. IMT_1 of cells where at least one daughter dies.
2. IMT_1 of cells where neither of the two daughters die.

Figure 3 shows the ECDFs of both groups. When comparing both curves with the Kolmogorov-Smirnov test, the p-value is 0.003, which provides statistical evidence against the null hypothesis: both datasets come from the same distribution. The IMT_1 distribution of cells in Group 1 tends to be delayed with respect to the IMT_1 distribution of Group 2. For example, 78% of cells in Group 2 have an IMT_1 below 22.5 hours, but only 46% of cells in Group 1 have an IMT_1 below this threshold. A similar effect, also statistically significant, is observed in the cells of the Third Generation (Figure 4):

78% of cells in Group 2 have an IMT_2 lower than 21 hours, but only 48% of cells in Group 1 have an IMT_2 below this threshold.

Now moving to the prediction of *Death* events in the Second Generation, we use the RIPPER algorithm with inputs DIV and IMT_1 to discover the following rule set:

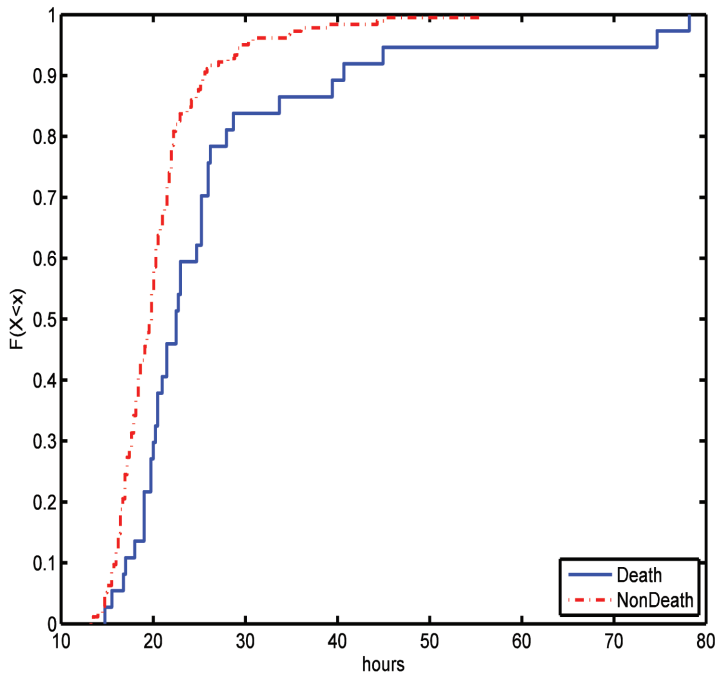
1. If ($IMT_1 \leq 18.75$ hours), no daughter cell dies.
2. Else if ($DIV \geq 21.5$ hours), no daughter cell dies.
3. Else if ($21.75 \leq IMT_1 \leq 22.25$) hours, no daughter cell dies.
4. Else if ($19.25 \leq IMT_1 \leq 21.25$) hours and ($DIV \leq 12.5$ hours), no daughter cell dies.
5. Else ($DIV \leq 4.5$ hours), no daughter cell dies.
6. Else at least one daughter cell dies.

(Note that to apply a RIPPER rule set to a database, the prediction of the first rule that fires is taken.) Using this rule set, the overall prediction accuracy is 75% (159 out of 213), where the accuracy for *Non-Death* cells is 78% (135 out of 176) and the accuracy for the other class is 65% (24 out of 37)¹. We use the WEKA (Hall et al., 2009) implementation of RIPPER (called JRIP) with default parameters². Because of high class imbalance, a cost-sensitive³ classifier is used, where zero cost is assigned to correct classifications and a cost inversely proportional to the size of the actual class is assigned to errors. Our aim is to require the classifier to learn both classes. We find a similar result when predicting Third Generation *Death* events. The rule set from JRIP is in this case:

Table 3. Cell death factor for Control lineages

	Gen. 1	Gen. 2	Gen. 3
Status= <i>Death</i>	13	51	40
Status= <i>Divided</i>	234	363	591
Cell death factor	5.5%	14%	6.7%

Figure 3. Comparison of the ECDFs of IMT_1 for cells that die and cells that do not



1. If ($IMT_1 \leq 22$ hours) and ($IMT_2 \leq 21.75$ hours), no daughter cell dies.
2. Else if ($22.75 \leq IMT_1 \leq 26.5$) hours, no daughter cell dies.
3. Else if ($IMT_2 \leq 17.75$ hours), no daughter cell dies.
4. Else if ($23.75 \leq IMT_2 \leq 26$) hours, no daughter cell dies.
5. Else at least one daughter cell dies.

The overall accuracy in this case is 80% (256 out of 320), while the accuracy for *Death* and *Non-Death* cell classes are respectively 67% (20 out of 30) and 81% (236 out of 290).

Study of the Cytotoxic Effect of Topotecan

The results are summarised in Table 4 where “Gen.” stands for Generation. It is statistically evident that the drug is killing more cells in the Perturbed lineages than in Controls in the First Generation, but there is no statisti-

cal evidence to say the same in the Second Generation. Surprisingly, the proportion of *Deaths* in the Second Generation for $1\mu\text{M}$ is lower than that for Control and $10\mu\text{M}$. For Stressed versus Resistant and Full Clonal Potential (FCP) versus Partial Clonal Potential (PCP), there is a trend from Control to $10\mu\text{M}$ in both generations: as more drug is added, there are more Stressed cells and fewer FCP cells. However, possibly due to the small sample size, we found the differences to be statistically significant only when comparing the Control and $10\mu\text{M}$ groups. The values of the proportions are shown in Table 5.

Cell Dynamic Behaviour. Here, we consider the dynamic behaviour of cells: that is, how the ECDFs of DIV , IMT_1 , IMT_2 and IMT_3 change when the system is perturbed with the drug. To compute the ECDF of each lineage type in each generation, the cells that have successfully divided in that generation are considered.

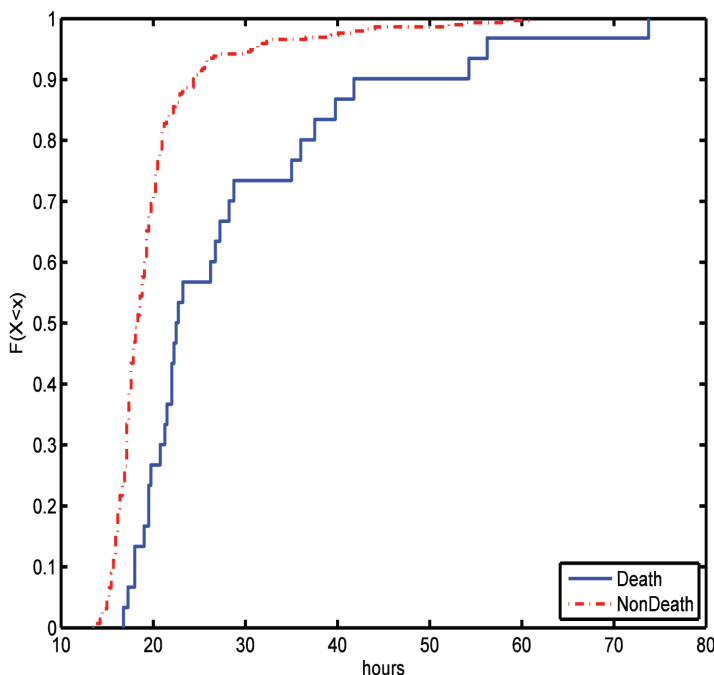
Figure 4. Comparison of the ECDFs of IMT_2 for cells that die and cells that do not

Figure 5 shows the ECDF of DIV for Control, $1\mu\text{M}$ and $10\mu\text{M}$ lineages. It shows a big discrepancy among the distributions of Control, $1\mu\text{M}$ and $10\mu\text{M}$ starting at 10 hours. $1\mu\text{M}$ is delayed and $10\mu\text{M}$ is greatly delayed when compared to Control. The Kolmogorov-Smirnov test finds statistical evidence *against* the null hypothesis for both Control versus $1\mu\text{M}$ ($p\text{-value} < 10^{-3}$) and Control versus $10\mu\text{M}$ ($p\text{-value} < 10^{-5}$). Just to illustrate this difference, 91% of the Control group's DIV times are less than 22 hours, but only 73% of the $1\mu\text{M}$ group's DIV and 50% of the $10\mu\text{M}$ group's DIV are below that threshold. The maximum difference between Control and $1\mu\text{M}$ ECDFs is 35%, which is at 19.25 hours; and between Control and $10\mu\text{M}$ ECDFs is 41%, which occurs at 22 hours.

Figure 6 shows the ECDF of IMT_1 for Control, $1\mu\text{M}$ and $10\mu\text{M}$ lineages. We observe that $1\mu\text{M}$ is slightly delayed with respect to Control, and $10\mu\text{M}$ is slightly delayed with

respect to $1\mu\text{M}$. The Kolmogorov-Smirnov test finds statistical evidence showing that the distribution of IMT_1 is different in Control and $10\mu\text{M}$ ($p\text{-value} < 10^{-5}$). 57% of Control IMT_1 are lower than 20.5 hours, but only 44% of $1\mu\text{M}$ IMT_1 and 18% of $10\mu\text{M}$ IMT_1 are below that threshold. The maximum difference between Control and $10\mu\text{M}$ is 39%, which is located at 20.5 hours.

Figure 7 shows the ECDFs of IMT_2 . We observe that $1\mu\text{M}$ is very similar to Control (it almost follows the Control distribution) and $10\mu\text{M}$ is delayed. The Kolmogorov-Smirnov test finds statistical evidence showing that the distribution of IMT_2 is different in Control group and $10\mu\text{M}$ group ($p\text{-value} < 10^{-5}$). There is also statistical evidence showing that the distribution of $1\mu\text{M}$ and $10\mu\text{M}$ are different ($p\text{-value} < 10^{-3}$). 60% of Controls and 58% of $1\mu\text{M}$ IMT_2 are less than 19.5 hours, but only 27% of $10\mu\text{M}$ IMT_2 are less than this threshold. The maximum difference between Controls and $10\mu\text{M}$ is 33%,

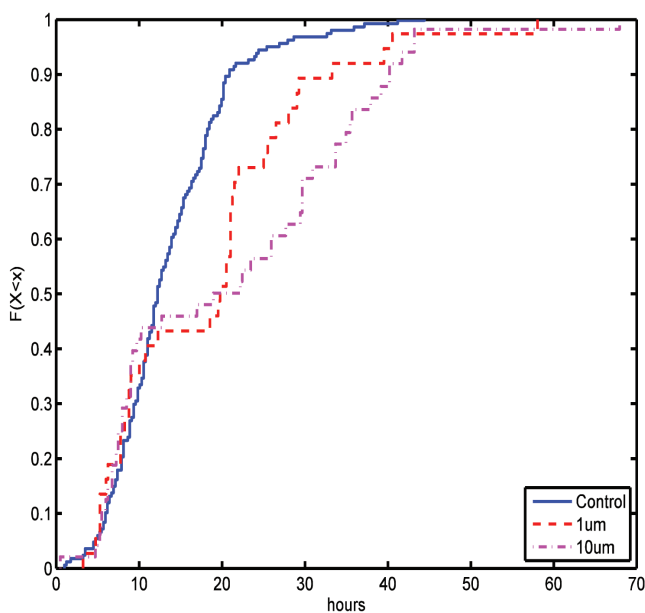
Table 4. Cytotoxic effect in the First and Second Generations for 1 μ M and 10 μ M. Those with statistically different proportions with respect to Control are shown in **boldface**. The values in parentheses are the corresponding p-values.

Proportion	Gen.	Control	1 μ M	10 μ M
Death/Non-Death	1 st	13/293	8/63 (0.038)	9/76 (0.033)
	2 nd	51/384	11/99	18/93
Stressed/Resistant	1 st	35/178	10/40	15/38
	2 nd	29/318	11/65	19/58 (<0.0002)
FCP/PCP	1 st	88/23	13/4	4/7 (<0.005)
	2 nd	196/13	30/3	11/6 (0.001)

Table 5. The values of proportions from Table 4

Proportion	Gen.	Control	1 μ M	10 μ M
Death/Non-Death	1 st	0.0444	0.1270	0.1184
	2 nd	0.1328	0.1111	0.1935
Stressed/Resistant	1 st	0.1966	0.2500	0.3947
	2 nd	0.0912	0.1692	0.3276
FCP/PCP	1 st	3.8261	3.2500	0.5714
	2 nd	15.0769	10.0000	1.8333

Figure 5. Comparison of the ECDFs of the Progenitor division time



which occurs at 19.5 hours; and between 1 μ M and 10 μ M is 32%, which is at 17.5 hours.

Figure 8 shows the ECDFs of IMT_3 . We observe the same trends as in Figure 7: the 1 μ M group is very close to the Control group and the 10 μ M group is closer to the Control group than before. The Kolmogorov-Smirnov test finds statistical evidence showing that the distribution of IMT_3 is different in Control and 10 μ M (p-value $<10^{-5}$). There is also evidence showing that the distribution of 1 μ M and 10 μ M are different (p-value $<10^{-4}$). 59% of 1 μ M IMT_3 and 57% of Control IMT_3 are lower than 19.5 hours, but only 29% of 10 μ M IMT_3 are lower than this threshold. The maximum difference between Control and 10 μ M is 28%, which is at 19.5 hours; and between 1 μ M and 10 μ M is 30%, which is at the same time point.

Study of Stressed Lineages

We now consider patterns that can be used to predict whether a lineage is Stressed or Resistant. We construct a series of classifiers using the following inputs:

1. Progenitor: Drug and DIV are used as input to the classifier.
2. First: Drug, DIV and First Generation information are used.
3. Second: Drug, DIV, First and Second Generations information are used.

Drug is the lineage type: Control, 1 μ M or 10 μ M.

The experiments are summarised in Table 6. The overall accuracy and the accuracy per class (Stressed and Resistant) are provided. The default parameters are used in WEKA's J48⁴ and JRIP. In this case, it is not necessary to handle class imbalance. We observe that the classifiers with the best accuracies are those that include Second Generation inputs. However, the classifiers using Second Generation information are not considered because the rules they provide are obvious and do not add more insight into lineages' behaviour. Therefore, the classifiers for the *First* input

data set are selected. The final rule set induced using JRIP for this input data is:

1. If ($Status_{1,1} = Arrested$), the lineage is Stressed.
2. Else if ($IMT_{1,2} \geq 39.5$ hours), the lineage is Stressed.
3. Else if ($Status_{1,2} = Death$), the lineage is Stressed.
4. Else if ($Status_{1,1} = Unresolved$), the lineage is Stressed.
5. Else the lineage is Resistant.

The first rule identifies that many of the *Arrested* cells in the First Generation have also an *Arrested* sister and therefore there are neither Second nor Third Generations in those lineages. The second rule shows that Stress lineages have at least one cell in the First Generation with a very long IMT_1 (the threshold is fixed to 39.5 hours). The next rule is obvious: if one cell's status is *Death*, then the other does not divide and the lineage must be stressed. The fourth rule is an artifact of the dataset: it happens in the data that when one cell in the First Generation is *Unresolved* the lineage is stressed.

The classifier resulting from the J48 algorithm is shown in Figure 9. This tree, which is also very simple, complements the JRIP ruleset for the Resistant lineages: they have at least one First Generation cell with an IMT lower than 36.5 hours.

Study of Control versus Perturbed Lineages

When we follow the procedure described in the previous section (training a series of classifiers with incremental generation information) to distinguish between Control and Perturbed lineages, the best result is given by J48 using all but Third Generation information. The overall accuracy of this classifier (Figure 10) is 78.4% (196 out of 250), where the accuracies on the Control and Perturbed lineages are 81.8% (135 out of 165) and 71.8% (61 out of 85) respectively. This classifier uses information of the time to

Figure 6. Comparison of the ECDFs of the first generation cells IMT

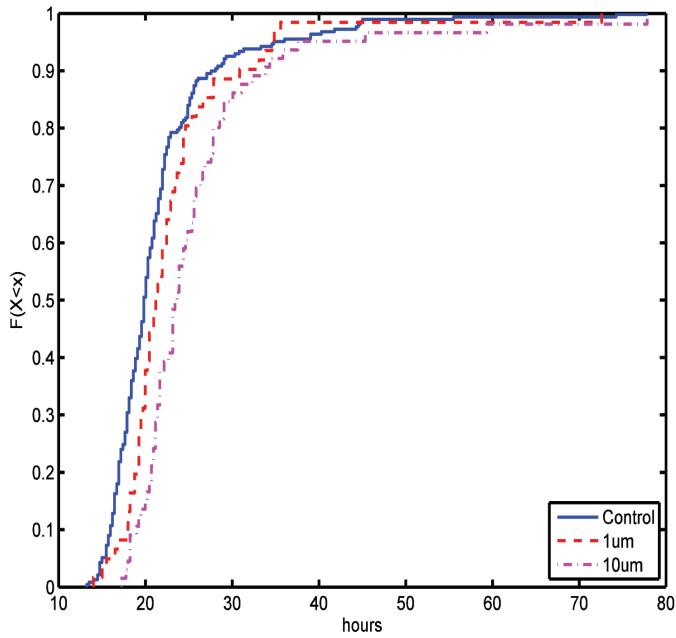
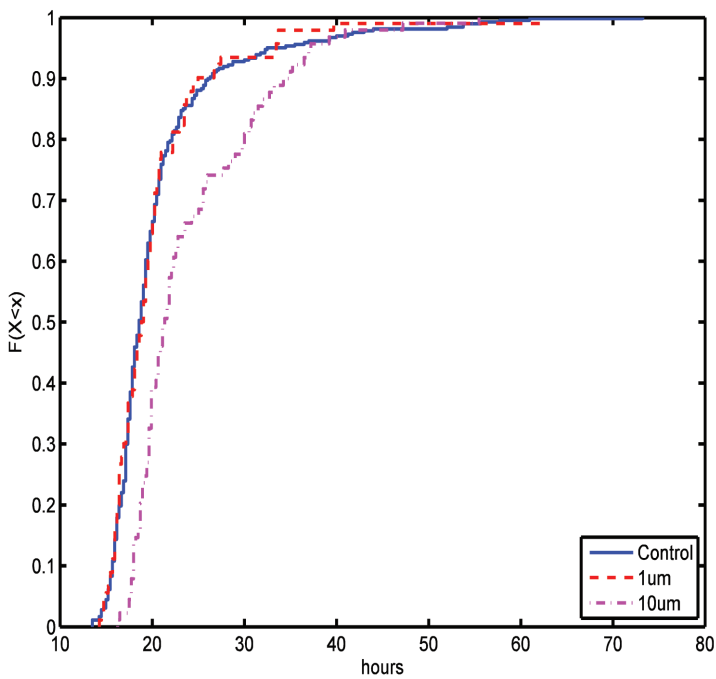


Figure 7. Comparison of the ECDFs of the second generation cells IMT



divide (DIV) and displacement (DIV_D) of the progenitor as its most discriminative features. For example, if a lineage progenitor cell takes longer than 20.7 hours to divide and moves more than 30.41 microns, the lineage is declared Perturbed. In addition uses $Cell_{1,2}$, $Cell_{2,4}$ and the number Second Generation Divisions to classify the lineages.

As it appears that using all the available information for training the classifier does not result in the best possible performance, next we explore Bayesian network modelling to find a suitable subset of variables.

Bayesian Network Modelling. In a Bayesian Network, the Markov Blanket of a variable T is the set that consist of its parents, its children and its children's parents (where these refer to the relationships between nodes in the Bayesian network, and are not to be confused with relationships between cells). If the values of the variables in the Markov Blanket are known, the rest of variables in the Bayesian Network become irrelevant for predicting the value of T .

Figure 11 shows the Markov Blanket of the best automatically learned network that we have induced from the data. This network is the result of applying to the data the supervised discretization proposed in Fayyad and Irani, (1993), followed by a local hill climbing search starting with the naive Bayes network. All those steps were done in WEKA. The overall accuracy of this network is 76% (190 out of 250), and the accuracy for Control and Perturbed classes are respectively 83% (137 out of 165) and 62.3% (53 out of 85).

The probability distribution tables provide further insights on the data. Table 7 shows that the events with a high probability for Control lineages have $Status_{1,2}=Divided$ with the exception of $Status_{3,5}=Arrested$. This probability reaches its maximum value (0.8) when both statuses are *Divided*. The events that show high probability for Perturbed lineages are $Status_{1,2}=Death$. In addition, the *Arrested* event in the First Generation is more associated with

Control lineages but the *Arrested* event in the Third Generation is more associated with Perturbed lineages.

Table 8 shows that the supervised discretization procedure selects a threshold for $IMT_{1,2}$ at 19.1 hours (which suggest that this threshold is informative for this task). We also observe that: (1) Control and Perturbed lineages behave differently when the $Cell_{2,4}$ cell is *Arrested*, as in the case of Control cells it is more probable that their mothers need less than 19.1 hours to divide but the opposite is more probable for Perturbed cells; (2) Control and Perturbed lineages behave similarly when a $Cell_{2,4}$ dies, as in both cases is very probable that its mother needs more than 19.1 hours to divide; (3) When the $Cell_{2,4}$ divides, there is a probability of 1/3 of $IMT_{1,2}$ being lower than 19.1 hours for Control lineages, but the probability of the same event is less than 1/12 for Perturbed lineages.

Table 9 shows that the supervised discretization selects a threshold for DIV at 9.4 hours and another one at 20.9 hours. We observe that most Control, as opposed to Perturbed, cells divide some time between 9.4 and 20.9 hours. We also observe that there is a positive correlation between DIV and DIV_D. Table 10 shows that the selected threshold for DIV_D is 59.4 microns. We observe that when the lineage is Control, the cell moves less than this threshold 91% of the time.

JRIP Result with Bayesian Variable Selection. If the data is assumed to be generated using a Bayesian network, all the relevant information about a variable is found in its Markov Blanket. Therefore, the Bayesian Network modeling can be used as a variable selection procedure, i.e., only the variables in the Markov Blanket of *Drug* are used to predict its outcome. Accordingly, we combine this variable selection procedure with a JRIP classifier (with both being performed inside the leave-one-out loop). Table 11 shows how many times each variable was selected. This method is consistent in all the leave-one-out itera-

Figure 8. Comparison of the ECDFs of the third generation cells IMT

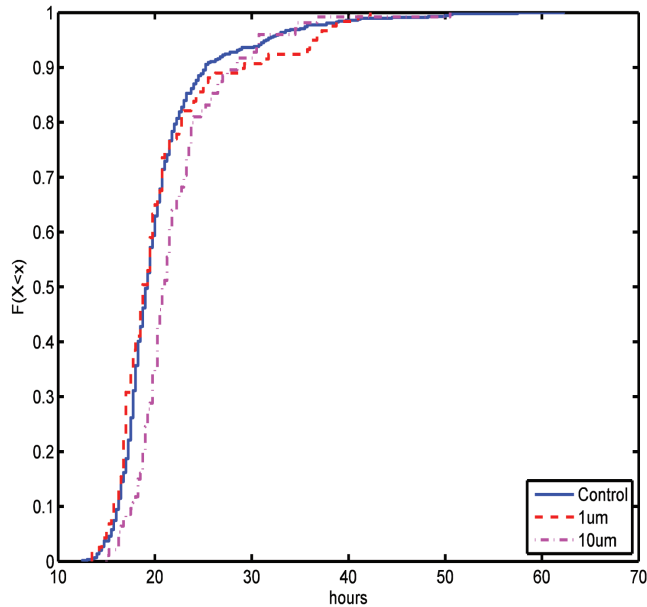


Table 6. Per-class and overall accuracies, leave-one-out result, for JRIP and J48 using inputs until generation “Gen.” (inclusive)

Gen.	Method	Overall	Stressed	Resistant
Progenitor	JRIP	76.0%	36.8%	90.0%
	J48	65.9%	0.0%	89.4%
First	JRIP	91.2%	75.4%	96.9%
	J48	92.2%	75.4%	98.1%
Second	JRIP	94.9%	89.5%	96.9%
	J48	95.8%	89.5%	98.1%

Figure 9. Tree classifier for the prediction of resistant versus stressed lineages using Progenitor and first generation information

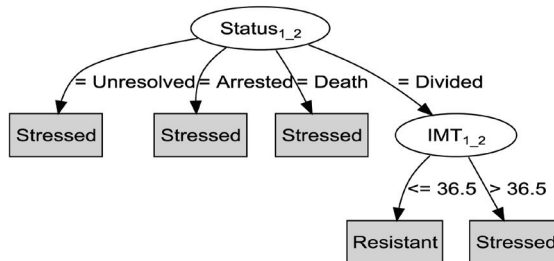


Figure 10. Tree classifier for the prediction of control versus perturbed lineages using Progenitor, first and second generations information

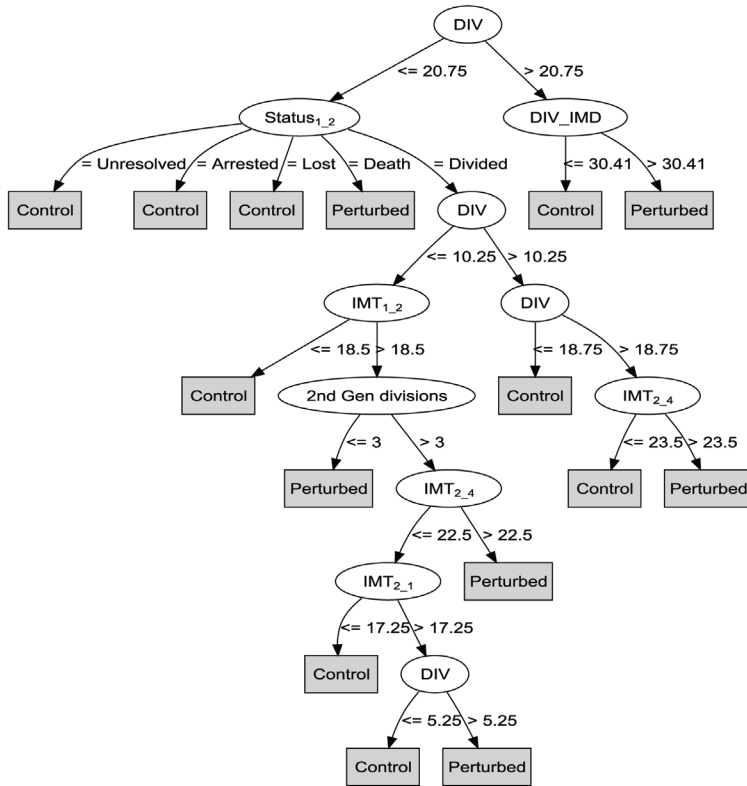


Figure 11. Markov Blanket of the automatically-learned Bayesian network for classifying perturbed versus control lineages (Drug)

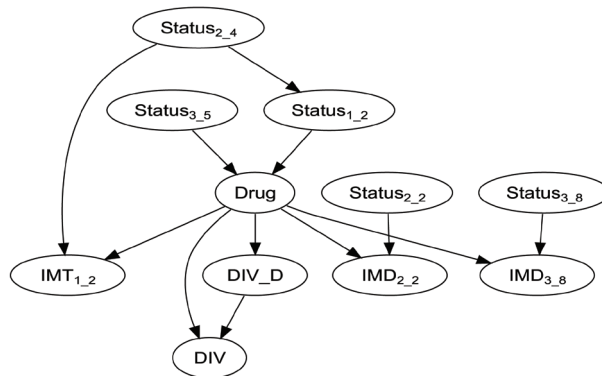


Table 7. Extract of the probability distribution of Drug given Status_{3,5} and Status_{1,2} for the network in Figure 11

Drug			
Status _{3,5}	Status _{1,2}	Control	Perturbed
Refused	Divided	75.0%	25.0%
Arrested	Divided	34.6%	65.4%
Lost	Lost	81.3%	18.8%
Lost	Divided	47.6%	52.4%
Never-Born	Unresolved	75.0%	25.0%
Never-Born	Arrested	87.9%	12.1%
Never-Born	Death	29.2%	70.8%
Never-Born	Divided	63.5%	36.5%
Death	Divided	59.1%	40.9%
Divided	Divided	80.4%	19.6%

Table 8. Extract of the probability distribution of IMT_{1,2} given Drug and Status_{2,4} for the network in Figure 11

IMT _{1,2}			
Drug	Status _{2,4}	(0, 19.1]	(19.1, ∞)
Control	Arrested	70.0%	30.0%
Control	Death	10.0%	90.0%
Control	Divided	32.9%	67.1%
Perturbed	Arrested	16.7%	83.3%
Perturbed	Death	5.0%	95.0%
Perturbed	Divided	7.8%	92.2%

Table 9. Probability distribution of DIV given drug and DIV_D for the network in Figure 11

DIV				
Drug	DIV_D	(0,9.4]	(9.4,20.9]	(20.9,inf)
Control	(0,59.4]	29.8%	60.0%	10.2%
Control	(59.4,inf)	3.2%	67.7%	29.0%
Perturbed	(0,59.4]	50.4%	16.2%	33.3%
Perturbed	(59.4,inf)	11.9%	15.3%	72.9%

Table 10. Probability distribution of DIV_D given drug for the network in Figure 11

Drug	DIV_D	
	(0,59.4]	(59.4,inf)
Control	91.3%	8.7%
Perturbed	66.9%	33.1%

tions: the same variables are selected in almost all the iterations. In some iterations, however, some other variables are included, but this happens less than 2% of the time.

This procedure's overall accuracy is 81% (203 out of 250), and the accuracy on Control and Perturbed classes are respectively 87.3% (144 out of 165) and 69.4% (59 out of 85). In this case, the JRIP rule set is:

1. If(DIV \geq 21 hours) and (DIV_D \geq 32.25 microns), the lineage is Perturbed.
2. Else if (DIV \leq 10 hours) and (IMT_{1,2} \geq 22.7 hours), the lineage is Perturbed.
3. Else the lineage is Control.

Performing variable selection using the Markov blanket in this way has yielded an effective and simple model of the data that has better accuracy than previous ones.

Dynamic Behaviour. IMD is the Euclidean distance from point where a cell is born to the point where a cell divides. This is not the same as the total cell displacement, because the cell can move in any direction and doesn't have to maintain one direction. However, we can assume that this movement can be characterised by an unknown probability distribution; as we have samples from this distribution, we can compare perturbed and unperturbed populations with the ECDF. The ECDFs of the 'displacement' of the cells (IMD) are now considered for Progenitor to Third Generation. These complement the IMT ECDFs that were shown in Figures 5 to 8.

Figure 12 shows the ECDFs of the displacement of Progenitor cells. We observe a clear difference between the distributions of the movements among Control, 1 μ M and 10 μ M groups, with Control cells having the lowest displacements and 10 μ M cells having the greatest displacements. For example, 77% Controls, 65% 1 μ M and 44% 10 μ M are lower than 37 microns. The maximum difference between Control and 1 μ M is 26%, which is located at 27 microns. The maximum difference between Control and 10 μ M is 33%, which is located at 37 microns. The differences between Control and 1 μ M and between Control and 10 μ M are statistically significant using a Kolmogorov-Smirnov test (p-values $<10^{-3}$). Comparing this figure with Figure 5, we observe that the ECDFs have different profiles; for example, 10 μ M and Controls are separated from the beginning of the graph in Figure 12.

Figure 13 shows the ECDFs of the First Generation cells IMD, the distributions are closer to each other and no statistical evidence is found for a difference between them. Again, these ECDFs have a different profile to the corresponding IMT ones in Figure 6.

Figure 14 shows the ECDFs of Second Generation cells IMD. The ECDFs are even closer to each other in this generation and 1 μ M cells are delayed more than 10 μ M cells over some ranges. We found no statistically significant differences between the curves using a Kolmogorov-Smirnov test. There is a notable difference between these ECDFs and its IMT: in Figure 7 the 1 μ M distribution is almost overlaid with the Control distribution, but this is not the case in its IMD counterpart.

Figure 15 shows the ECDFs of Third Generation cells IMD: the distributions of 1 μ M

Table 11. Bayesian variable selection in each leave-one-out iteration. The number of times a variable is selected by the procedure is recorded

Variable	number of times selected
DIV	250
IMT _{1,1}	5
IMT _{1,2}	250
DIV_D	250
IMD _{2,2}	250
IMD _{3,8}	250
Status _{1,2}	250
Status _{2,2}	249
Status _{2,4}	243
Status _{3,4}	1
Status _{3,5}	250
Status _{3,8}	250

and 10 μ M IMD are, surprisingly, more delayed than in previous generations with respect to Control, breaking the trend that was observed in the previous generations. For example, 71% Controls, 52% 1 μ M and 44% 10 μ M are lower than 51 microns. The maximum difference between Control and 1 μ M is 20% and it is located at 55 microns. The maximum difference between Control and 10 μ M is 27% and it is located at 51 microns. Both differences are statistically significant using a Kolmogorov-Smirnov test (p-values $<10^{-3}$). As before, we observe that the behaviour is different to that of the corresponding IMT graph (Figure 8).

Cell ‘Speed’ and cytotoxic level. The previous subsection shows that the distribution of IMD are different for Control and Perturbed populations. Here we show that the relative ‘speed’ distributions of the cell are also different. We define the *speed* of any particular cell as the ratio of its IMD to its IMT. Figures 16 to 19 show the empirical distribution of this ratio.

Figure 16 and Figure 17 show this difference for Progenitor and First Generation cells. We observe that 10 μ M moves slightly faster than Controls, but we do not find this difference to be statistically significant with a Kolmogorov-Smirnov test.

Figure 18 shows this difference for Second Generation cells. Again, we observe that there appear to be slight differences between the distributions, but these are not found to be statistically significant with a Kolmogorov-Smirnov test.

Figure 19 shows this difference for Third Generation cells. We observe that 1 μ M moves slightly faster than Controls and 10 μ M moves faster than both. This time, both differences are statistically significant with a Kolmogorov-Smirnov test when compared to Control (p-values $<3 \cdot 10^{-4}$). For example, 73% Controls, 54% 1 μ M and 49% 10 μ M are lower than 2.7 micron/hour. The maximum difference between Control and 1 μ M is 21% and it is located at 2.8 micron/hour. The maximum difference between Control and 10 μ M is 24% and it is located at 2.7 micron/hour.

DISCUSSION

Here we analyse the results and propose some hypotheses that might be worthy to explore further.

Study of Cell Death in Control Lineages

Cell death, is deemed to be a part of the regulatory mechanism of tumour system, which invokes detailed investigation of such dynamic system. This work has revealed the following patterns for the *Death* event for Control lineages: Firstly, there is a strong pattern of paired sister death, i.e. when one cell dies, it is very probable that its sister also dies; Secondly, cell death in general can be related with longer cell cycle duration of the corresponding mother cell.

Study of the Cytotoxic Effect of Topotecan

The cytotoxic effect of the drug is prominent in 10 μ M, where statistically significant differences between Control and 10 μ M lineages are found in almost all tests. On the other hand, the effect of the drug in 1 μ M has only a statistical significant impact in the *Death to Non-Death* proportion and in the ECDFs of DIV and IMT₁. *Death Non-Death* proportion in the Second Generation is lower for 1 μ M than both Control and 10 μ M. This might be an artifact of the data or something to be further investigated: *does a low dosage of drug result in fewer cells death the Second Generation after the drug administration is stopped?* The trends in the other cases suggest that in low dose, i.e., 1 μ M has still some cytotoxic effect, but it is almost lost in the Second Generation, as we see in Figure 7 where 1 μ M IMT₂ distribution is very close to IMT₂ Control distribution (although its IMD is still affected as shown in Figure 14). The greatest variation between Control and 10 μ M is found in the proportion of Full Clonal Potential to Partial Clonal Potential. 10 μ M drug dosage makes this proportion 6.7 times lower than Control for the First Generation and 8.2 times lower for the Second Generation.

There is an interesting relationship found in the ECDF of DIV (Figure 5): For progenitor cells, TPT prevents division between the times T=10 hours and T=20 hours, establishing the S-phase specific cytotoxic effect of TPT (Pommier, 2006). When TPT was administered, the progenitor cells were distributed in all 3 phases (G₂, S and G₂M) phases of the cell cycle, and from the start of the experiment the cohorts of progenitor cells that delivers 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 occurred after 18 h. (Feeney et al., 2003; Errington, Marquez, Chappell, Wiltshire, & Smith, 2005).

Study of Stressed Lineages

In the algorithms to classify the lineages (Control population also has stressed lineages), the *Drug* variable is not selected, which suggests, as the accuracy of the classifiers is quite high, that we do not learn much from the *Drug* variable after considering the other data. The key factor that explains whether a lineage will be stressed is the behaviour of the lineage in First Generation, for all lineage types. In addition, *Arrested* cells tend to appear as sister pairs in the First Generation. For example, for Control lineages, 48 out of 49 *Arrested* cells come from *Arrested* sister pairs.

Study of Control versus Perturbed Lineages

When we analyse a data set with a small number of examples relative to the number of variables, we might expect to encounter the well-known *curse of dimensionality*. In general, as the number of variables increases, the number of possible hypotheses to explain the data increases (assuming a sufficiently rich representation language), which increases the risk of discovering hypotheses that are based just on artefacts of the data, which in turn leads to models having lower accuracies. In addition, correlated variables, noisy variables and partially redundant variables all increase

Figure 12. Comparison of the ECDFs of the Progenitor displacement

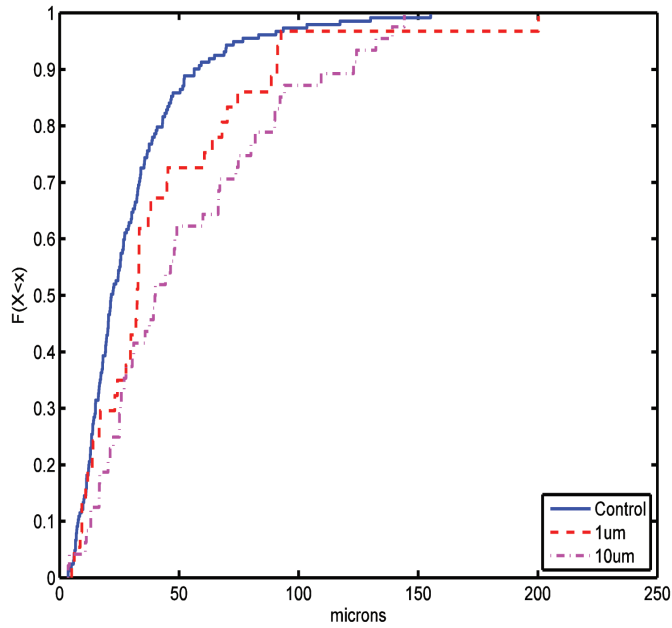
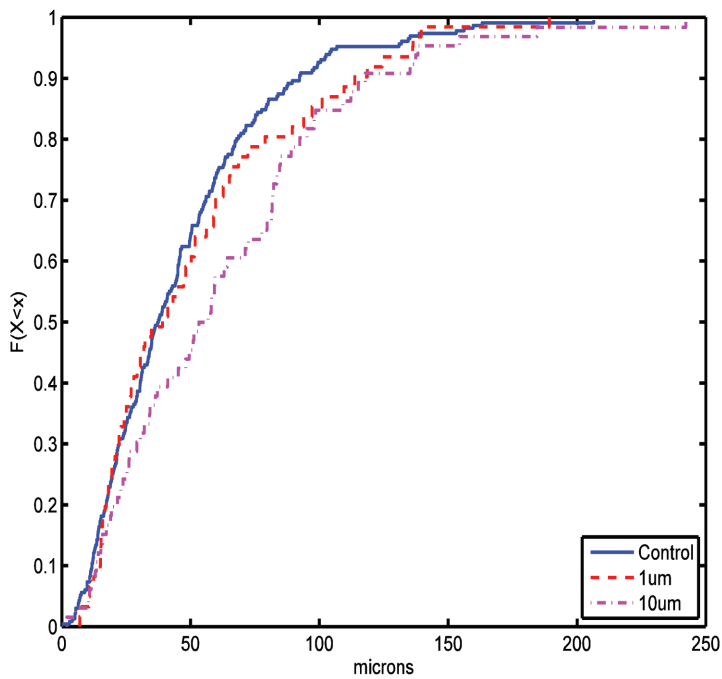


Figure 13. Comparison of the ECDFs of the first generation cells IMD



the hypotheses search space, making the search harder, without providing new information. We suggest that this is the reason why we get models with better accuracy when we follow the common practice of feature selection, rather than using all the variables.

We hypothesise that the extra displacement of $1\mu\text{M}$ and $10\mu\text{M}$ lineages, when compared to Control lineages, may be caused by some internal cell mechanism that compels the cell to move away from stressful environments. If we look at Figure 12, we see that the greater the concentration of drug, the greater the displacement, which gives some support to this hypothesis. If we look at the movement of the cells, the general trend is that Control cells move less than $1\mu\text{M}$ cells, which in turn move less than $10\mu\text{M}$ cells. This does not happen in the Second Generation, where $1\mu\text{M}$ and $10\mu\text{M}$ distributions are interlaced, and over some ranges the $10\mu\text{M}$ distribution is closer to the Control distribution (see Figure 14).

When comparing IMD and IMT distributions, we observe that in the Second Generation, the $1\mu\text{M}$ IMT ECDF is very close to Control distribution, but the same does not happen with the $1\mu\text{M}$ IMD ECDF. This is also found in the Third Generation. The greatest differences among IMD ECDFs are observed in Progenitor and Third Generation cells.

When comparing the cell *speed* distributions we observe that the difference among populations is not only in the distance they travel but also in the speed at they move, which indicates that perturbed populations not only move more but also faster. As a general trend, $10\mu\text{M}$ is the fastest population with the exception of Second Generation (Figure 18) where $1\mu\text{M}$ is the fastest one. This is related to what we already noted above (see Figure 14) where $1\mu\text{M}$ and $10\mu\text{M}$ IMD distributions are similar, but we learn from Figure 18 that $1\mu\text{M}$ cells are moving faster in a shorter IMT. Also surprising is that the behaviour of $1\mu\text{M}$ departs from that of the Control group in Second Generation. The behaviour of both $1\mu\text{M}$ and $10\mu\text{M}$ is clearly different (statistically significant) from Control in Third Generation (Figure 19).

From these observations, we conclude that there is something more than a simple correlation between IMT and IMD. This leads to the following new questions and hypothesis to explore in the future: Why do Perturbed cells move more? Does the drug in small dosages make cancer cells expand more rapidly?

The informative events (probabilities close to zero or one) in Bayesian networks probability tables help us to consider interesting expected and unexpected behaviours. In this case we focus on the *Arrested* event: cells exhibit different behaviours in Control and Perturbed lineages as expected. However, it is surprising how the probability of the event changes from the First to the Third Generation and how it changes depending on the time taken by the mother to divide. We also observed that the division time of the ancestors has an impact on the probability of some future events of their offspring.

By modelling the data with a Bayesian Network, we can perform a variable selection that allows us to subsequently discover simpler and better models using rule-based learners. The JRIP model gets to 81% overall accuracy, it is simple and uses the displacement information. Again, cell movement is important and increases accuracy.

CONCLUSION

Through this work, we have demonstrated how data mining procedures can reveal important patterns of a dynamic cell system. By our analyses, we have found the cytotoxic effect of Topotecan on both the dynamics and event frequencies of the system. The results reiterate the finding that initial cell cycle positioning is an important factor for immediate cytotoxic response. The results also suggest that by distributing the cytotoxic effect asymmetrically within the progeny, the system adopts strategies that facilitate the generation of drug resistance progenies, thus maximizing its clonal expansion potential.

We also observe cellular deaths in unperturbed condition with well defined patterns,

Figure 14. Comparison of the ECDFs of the second generation cells IMD

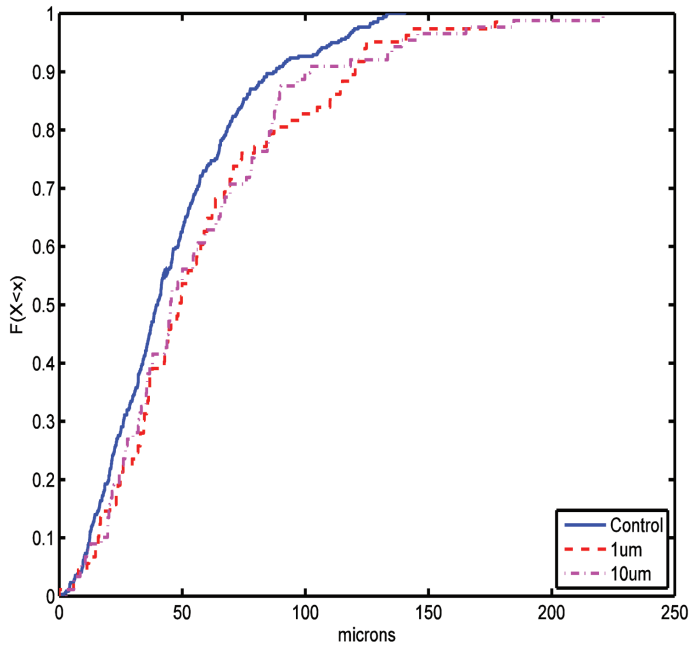


Figure 15. Comparison of the ECDFs of the third generation cells IMD

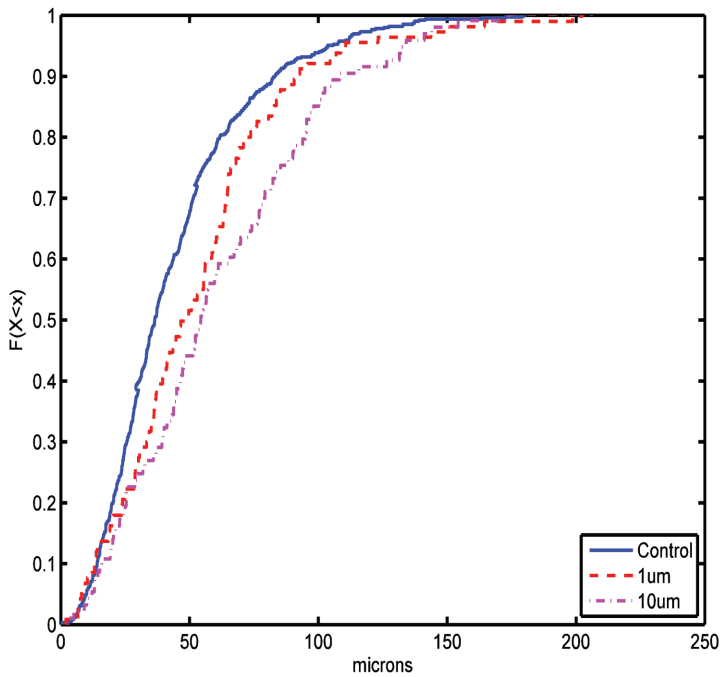


Figure 16. Comparison of the ECDFs of the Progenitor displacement – time to divide ratio

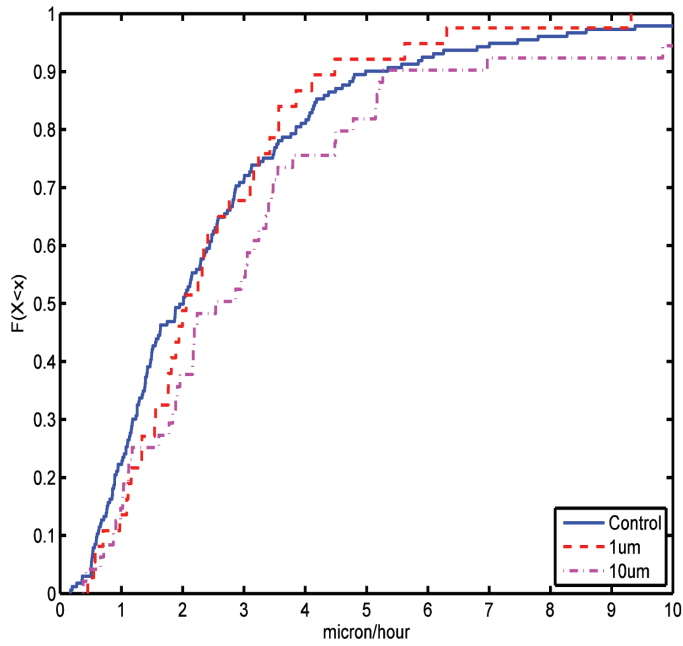


Figure 17. Comparison of the ECDFs of the first generation cells IMD – IMT ratio

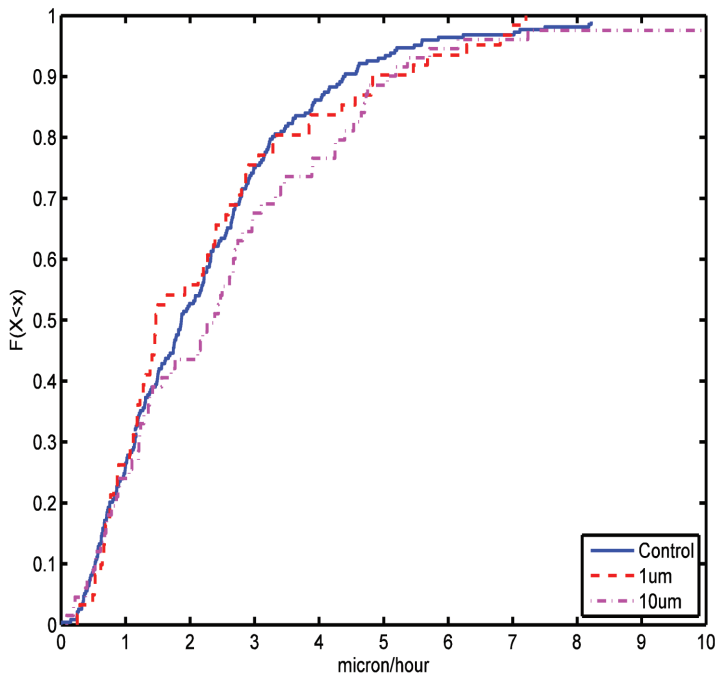


Figure 18. Comparison of the ECDFs of the second generation cells IMD – IMT ratio

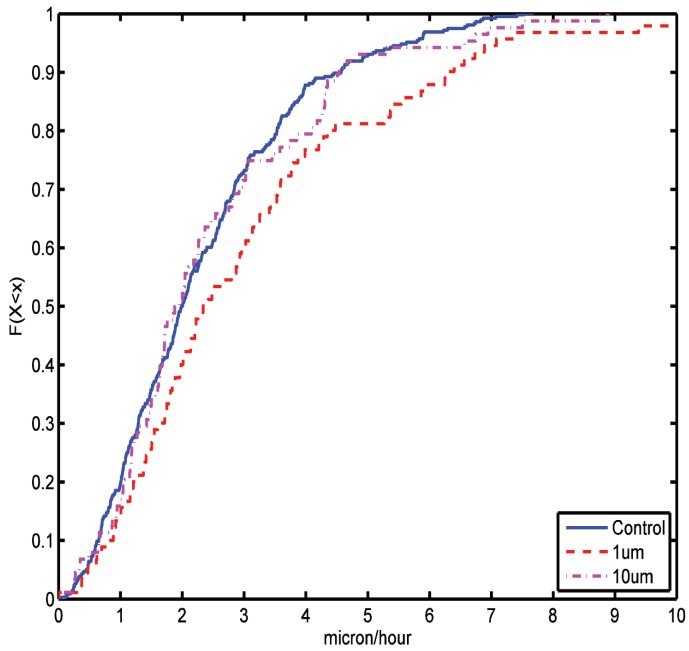
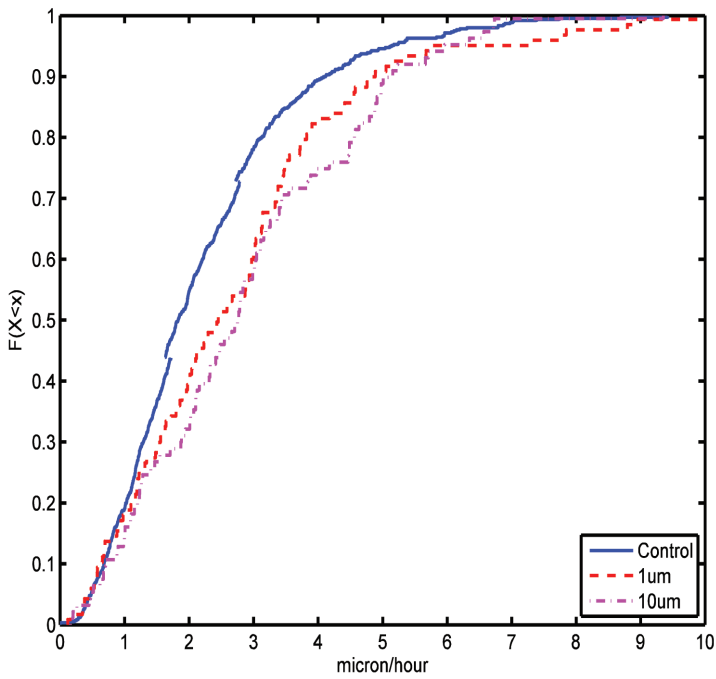


Figure 19. Comparison of the ECDFs of the third generation cells IMD – IMT ratio



indicating the existence of an innate regulation mechanism of the tumour system. Investigation on such a mechanism requires incorporation of gene and protein network data. In addition, we have found simple models that can predict death in cells in an unperturbed condition, and also how inheritance information like the cell cycle duration of a mother cell can influence the fate of daughters. Our models are successful at classifying between perturbed and unperturbed lineages, thus revealing the population level cytotoxic effect. More importantly, by predicting the clonal potential, we are not only able to identify the resistant subpopulation in early generations, but we are also able to identify strategies that the following generations adopt in order to maximize their clonal potential. In the spatial context, we have observed a putative relationship between cytotoxic effect and cellular movement; such results can be exploited to understand the metastasis potential of the tumour (Cavanna, Pokorna, Vesely, Gray, & Zicha, 2007), which is another important aspect of cancer research.

ACKNOWLEDGMENTS

This research has been supported by a Marie Curie Transfer of Knowledge Fellowship of the EU 6th Framework Programme, contract CT-2005-029611. It has also been supported in part by the Science Foundation Ireland under Grant 08/RFP/CMS1254. Ricardo Santiago-Mozos has been partly supported by Fundación Española para la Ciencia y la Tecnología (Ministerio de Educación de España).

REFERENCES

- Alvarez-Buylla, A., García-Verdugo, J. M., & Tramontin, A. D. (2001). A unified hypothesis on the lineage of neural stem cells. *Nature Reviews. Neuroscience*, 2, 287–293. doi:10.1038/35067582
- Anderson, D. J., Gage, F. H., & Weissman, I. L. (2001). Can stem cells cross lineage boundaries? *Nature Medicine*, 7, 393–395. doi:10.1038/86439
- Ardavín, C., Martínez del Hoyo, G., Martín, P., Anjuère, F., Arias, C. F., & Marín, A. R. (2001). Origin and differentiation of dendritic cells. *Trends in Immunology*, 22, 691–700. doi:10.1016/S1471-4906(01)02059-2
- Bailly, C. (2000). Topoisomerase I poisons and suppressors as anticancer drugs. *Current Medicinal Chemistry*, 7, 39–58. doi:10.2174/0929867003375489
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B. Methodological*, 57, 289–300.
- Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29, 1165–1188. doi:10.1214/aos/1013699998
- Bernards, R., & Weinberg, R. A. (2002). Metastasis genes: A progression puzzle. *Nature*, 418, 823. doi:10.1038/418823a
- Campbell, I. (2007). Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. *Statistics in Medicine*, 26, 3661–3675. doi:10.1002/sim.2832
- Carlson, J. M., Heckerman, D., & Shani, G. (2009). *Estimating false discovery rates for contingency tables* (Tech Rep. No. 2009-53). Microsoft Research.
- Cavanna, T., Pokorna, E., Vesely, P., Gray, C., & Zicha, D. (2007). Evidence for protein 4.1 B acting as a metastasis suppressor. *Journal of Cell Science*, 120, 606–616. doi:10.1242/jcs.000273
- Charniak, E. (1991). Bayesian Networks without Tears. *AI Magazine*, 12, 50–63.
- Cheung, S., Evans, N. D., Chappell, M. J., Godfrey, K. R., Smith, P. J., & Errington, R. J. (2008). Exploration of the intercellular heterogeneity of topotecan uptake into human breast cancer cells through compartmental modelling. *Mathematical Biosciences*, 213, 119–134. doi:10.1016/j.mbs.2008.03.008
- Cohen, W. W. (1995). Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning* (pp. 115–123).
- Dor, Y., Brown, J., Martinez, O. I., & Melton, D. A. (2004). Adult pancreatic beta-cells are formed by self-duplication rather than stem-cell differentiation. *Nature*, 429, 41–46. doi:10.1038/nature02520

- Errington, R. J., Marquez, N., Chappell, S. C., Wiltshire, M., & Smith, P. J. (2005). Time-Lapse Microscopy Approaches to Track Cell Cycle Progression at the Single-Cell Level. In *Current Protocols in Cytometry*. New York: John Wiley & Sons. doi:10.1002/0471142956.cy1204s31
- Fayyad, U. M., & Irani, K. B. (1993). Multi-interval discretization of continuous valued attributes for classification learning. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence* (Vol. 2, pp. 1022-1027). San Francisco: Morgan Kaufmann Publishers.
- Feeney, G. P., Errington, R. J., Wiltshire, M., Marquez, N., Chappell, S. C., & Smith, P. J. (2003). Tracking the cell cycle origins for escape from topotecan action by breast cancer cells. *British Journal of Cancer*, 88, 1310–1317. doi:10.1038/sj.bjc.6600889
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11, 10–18. doi:10.1145/1656274.1656278
- Heppner, G. H. (1984). Tumor heterogeneity. *Cancer Research*, 44, 2259–2265.
- Hope, K. J., Jin, L., & Dick, J. E. (2004). Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. *Nature Immunology*, 5, 738–743. doi:10.1038/ni1080
- Howard, R. A., & Matheson, J. E. (2005). Influence diagrams. *Decision Analysis*, 2, 144–147. doi:10.1287/deca.1050.0050
- Ichinose, Y., Seto, T., Nishiwaki, Y., Kiura, K., Sakai, H., & Yokoyama, A. (2010). Phase I Study of Topotecan and Cisplatin in Patients with Small Cell Lung Cancer. *Japanese Journal of Clinical Oncology*. doi:10.1093/jjco/hyq177
- Khan, I. A., Hedley, C. J., White, N. S., Ali, R., Chappell, M. J., & Evans, N. D. (2006). A novel integrative bioinformatics environment for encoding and interrogating timelapse microscopy images. In *Modelling and Control in Biomedical Systems* (pp. 273–278). Amsterdam, The Netherlands: Elsevier.
- Khan, I. A., Husemann, P., Campbell, L., White, N. S., White, R. J., & Smith, P. J. (2007). ProgeniDB: a novel cell lineage database for generation associated phenotypic behavior in cell-based assays. *Cell Cycle (Georgetown, Tex.)*, 6, 868–874. doi:10.4161/cc.6.7.4045
- Kim, K. M., & Shibata, D. (2002). Methylation reveals a niche: stem cell succession in human colon crypts. *Oncogene*, 21, 5441–5449. doi:10.1038/sj.onc.1205604
- Lorusso, D., Pietragalla, A., Mainenti, S., Masciullo, V., Di Vagno, G., & Scambia, G. (2010). Review role of topotecan in gynaecological cancers: current indications and perspectives. *Critical Reviews in Oncology/Hematology*, 74, 163–174. doi:10.1016/j.critrevonc.2009.08.001
- Noctor, S. C., Flint, A. C., Weissman, T. A., Dammerman, R. S., & Kriegstein, A. R. (2001). Neurons derived from radial glial cells establish radial units in neocortex. *Nature*, 409, 714–720. doi:10.1038/35055553
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Francisco: Morgan Kaufmann.
- Pommier, Y. (2006). Topoisomerase I inhibitors: camptothecins and beyond. *Nature Reviews. Cancer*, 6, 789–802. doi:10.1038/nrc1977
- Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4, 77–90.
- Rubin, H. (1990). The significance of biological heterogeneity. *Cancer and Metastasis Reviews*, 9, 1–20. doi:10.1007/BF00047585
- Seibel, N. L., Krailo, M., Chen, Z., Healey, J., Breitfeld, P. P., & Drachtman, R. (2007). Upfront window trial of topotecan in previously untreated children and adolescents with poor prognosis metastatic osteosarcoma: childrens Cancer Group (CCG) 7943. *Cancer*, 109, 1646–1653. doi:10.1002/cncr.22553
- Smirnov, N. (1939). Estimate of deviation between empirical distribution functions in two independent samples. *Bull. Moscow Univ*, 2, 3–16.
- Smith, P. J., Khan, I. A., & Errington, R. J. (2009). Cytomics and cellular informatics - coping with asymmetry and heterogeneity in biological systems. *Drug Discovery Today*, 14, 271–277. doi:10.1016/j.drudis.2008.11.012
- Stern, C. D., & Fraser, S. E. (2001). Tracing the lineage of tracing cell lineages. *Nature Cell Biology*, 3, E216–E218. doi:10.1038/ncb0901-e216

Tang, M., Pires, Y., Schultz, M., Duarte, I., Gallegos, M., & Wistuba, I. I. (2003). Microsatellite analysis of synchronous and metachronous tumors: a tool for double primary tumor and metastasis assessment. *Diagnostic Molecular Pathology*, 12, 151–159.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Wang, J. C. (1996). DNA topoisomerases. *Annual Review of Biochemistry*, 65, 635–692. doi:10.1146/annurev.bi.65.070196.003223

Weigelt, B., Glas, A. M., Wessels, L. F., Witteveen, A. T., Peterse, J. L., & Vant Veer, L. J. (2003). Gene expression profiles of primary breast tumors maintained in distant metastases. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 100, pp. 15901-15905). doi:10.1073/pnas.2634067100

Yamamoto, N., Yang, M., Jiang, P., Xu, M., Tsuchiya, H., & Tomita, K. (2003). Determination of clonality of metastasis by cell-specific color-coded fluorescent-protein imaging. *Cancer Research*, 63, 7785.

ENDNOTES

- 1 The 51 cell deaths in the Second Generation are comprised of 14 (*Death, Death*) sister pairs and 23 (*Death, non-Death*) sister pairs.
- 2 Default parameters for JRIP are: 'error rate $\geq \frac{1}{2}$ ' is included in stopping criterion; 1 of 3 folds is used for pruning; the minimum total weight of the instances in a rule is set to 2; and the initial seed is 1. We perform 25 optimization runs to minimise the sensitivity to the initial seed.
- 3 The cost-sensitive classifier, also implemented in WEKA (CostSensitiveClassifier), uses the resampling method.
- 4 Default parameters for the J48 are: not binary splits on nominal attributes; confidence factor used for pruning equals 0.25; minimum number of instances per leaf equals 2; 1 of 3 folds using for pruning, the rest for growing the tree; C.4.5 pruning; sub-tree raising operation when pruning.

Ricardo Santiago-Mozos is a Spanish Foundation for Science and Technology post-doctoral fellow (since September 2010) at the Machine Learning & Data Mining Group in National University of Ireland, Galway where he has just completed a Marie Curie Fellowship that commenced in October 2009. In September 2009 he obtained his PhD in Signal Processing and Communications at the Carlos III University of Madrid, where he developed a screening system for tuberculosis diagnosis. His research interests include machine learning, medical applications, genetic algorithms and communications, and he has co-authored more than 20 papers appearing in refereed journals and conference proceedings.

Imtiaz A. Khan (IAK) has been active in the fields of bioinformatics and computational biology and has focused lately (over 5 yrs) on image based oncology research. IAK started his professional career in pharmaceutical sciences; and then actively converted to gaining training in bioinformatics inspired by the challenges facing biology and medicine in embracing bioinformatics science and technology. IAK gained post-graduate degrees in Bioinformatics from UK and US institutions along with associated training elements at both NCBI and EBI. IAK has introduced novel bioinformatics infrastructure (PMID: 17387278) that aims to transform images to knowledge and thus augmenting our understanding about the dynamics of cellular behaviour. He is a senior researcher within the Cardiff-Swansea Systems Cytometry Group, an interdisciplinary group bringing together biologists, engineers and mathematicians alike. Currently, as a Marie Curie Fellow he is working at the Broad Institute of Harvard and MIT, Cambridge, USA. Primary objective of his current research is to develop infrastructure enabling the interoperability of data generated from different cytometric platforms.

Michael G. Madden is a Senior Lecturer in the College of Engineering and Informatics in the National University of Ireland Galway, which he joined in 2000. In 2006/07, he spent a year as a Visiting Research Scientist in the University of Helsinki, the University of California Irvine, and the University of California Berkeley. He previously worked in professional software R&D. He established and leads the Machine Learning & Data Mining Group in NUI Galway (<http://datamining.it.nuigalway.ie>). His inter-disciplinary research focuses on the development of new machine learning methods (Bayesian, SVMs, and others) and their application to scientific, medical and other domains, and includes contributions in theoretical algorithms and practical applications. He has published over 60 papers and has received seven publication awards. He has three patent filings, and he co-founded a spin-out company, Analyze IQ Limited, based on research that he led.