Table of Contents

IEEE ICDM Workshop on Biological Data Mining and its Applications in Healthcare

Editorial Preface

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 descendent 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 progenitor 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, Dammberner, & Kriegstein, 2001) and medicine (Bernards & Weinberg, 2002; Hope, Jin, & Dick, 2004; Tang et al., 2003; Weigel 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 tumor 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 McCoys 5A modified (Sigma) supplemented with 2 mM glutamine, 100 units/ml penicillin, 100 mg/ml streptomycin, 10% fetal calf serum and 1000 mg/ml genetin.

Cells are treated with 1 µM and 10 µM bolus dose of anticancer agent Topotecan (TPT) (Bailly, 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 1µM and 48 are 10µM 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, Cell₁₁ and Cell₁₂ in Figure 2 are sisters; both are daughters of Cell₁₁, which is a daughter of Cell₀₁, which is again a daughter of a Progenitor cell. The daughters of a Progenitor are termed First Generation cells (Cell₁₁ and Cell₁₂ in Figure 2). The daughters of the First Generation cells are termed Second Generation cells (Cell₂₁, …, Cell₂₄ 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 Cell₁₁ and the other as Cell₁₂. The daughters of Cell₁₁ are called Cell₁₂₁ and Cell₁₂₂, and are sorted according their
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\_I\_J (14 variables, one for each cell apart from the Progenitor cell; see Table 2 for their possible values).
3. IMT\_I\_J and IMD\_I\_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 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](image-url)
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 1. Notation and definitions**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell&lt;sub&gt;1&lt;/sub&gt; &lt;sub&gt;J&lt;/sub&gt;</td>
<td>Cell J in Generation I.</td>
</tr>
<tr>
<td>IMT</td>
<td>Inter-Mitotic Time. The time needed by a cell to divide.</td>
</tr>
<tr>
<td>IMD</td>
<td>Inter-Mitotic Displacement. The distance travelled by a cell until its division.</td>
</tr>
<tr>
<td>DIV</td>
<td>The elapsed time from the start of the experiment until division of a Progenitor cell.</td>
</tr>
<tr>
<td>DIV&lt;sub&gt;D&lt;/sub&gt;</td>
<td>The distance travelled by a Progenitor cell from the start of the experiment until division.</td>
</tr>
<tr>
<td>Status&lt;sub&gt;1&lt;/sub&gt; &lt;sub&gt;J&lt;/sub&gt;</td>
<td>Status of Cell J in Generation I.</td>
</tr>
<tr>
<td>IMT&lt;sub&gt;1&lt;/sub&gt; &lt;sub&gt;J&lt;/sub&gt;</td>
<td>The IMT for cell J in Generation I.</td>
</tr>
<tr>
<td>IMD&lt;sub&gt;1&lt;/sub&gt; &lt;sub&gt;J&lt;/sub&gt;</td>
<td>The IMD for cell J in Generation I.</td>
</tr>
<tr>
<td>Resistant</td>
<td>Cells whose descendants reach Fourth Generation. When applied to lineages, lineages that reach Fourth Generation.</td>
</tr>
<tr>
<td>Stressed</td>
<td>The opposite to resistant.</td>
</tr>
<tr>
<td>Control</td>
<td>Cells (lineages) that have not been exposed to drug.</td>
</tr>
<tr>
<td>Perturbed</td>
<td>Cells (lineages) that have been exposed to drug.</td>
</tr>
<tr>
<td>Drug</td>
<td>A variable that carries information to distinguish the type of lineage.</td>
</tr>
</tbody>
</table>
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, 1µM and 10µM 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 µM and 10 µM 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 et al. (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 distribution 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\textsubscript{1} (both IMT\textsubscript{1,1} and IMT\textsubscript{1,2}) is divided into two groups:

1. IMT\textsubscript{1} of cells where at least one daughter dies.
2. IMT\textsubscript{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\textsubscript{1} distribution of cells in Group 1 tends to be delayed with respect to the IMT\textsubscript{1} distribution of Group 2. For example, 78% of cells in Group 2 have an IMT\textsubscript{1} below 22.5 hours, but only 46% of cells in Group 1 have an IMT\textsubscript{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\textsubscript{2} lower than 21 hours, but only 48% of cells in Group 1 have an IMT\textsubscript{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\textsubscript{1} to discover the following rule set:

1. If (IMT\textsubscript{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\textsubscript{1} \leq 22.25) hours, no daughter cell dies.
4. Else if (19.25 \leq IMT\textsubscript{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). We use the WEKA (Hall et al., 2009) implementation of RIPPER (called JRIP) with default parameters. Because of high class imbalance, a cost-sensitive 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

<table>
<thead>
<tr>
<th>Status</th>
<th>Gen. 1</th>
<th>Gen. 2</th>
<th>Gen. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death</td>
<td>13</td>
<td>51</td>
<td>40</td>
</tr>
<tr>
<td>Divided</td>
<td>234</td>
<td>363</td>
<td>591</td>
</tr>
<tr>
<td>Cell death factor</td>
<td>5.5%</td>
<td>14%</td>
<td>6.7%</td>
</tr>
</tbody>
</table>
1. If \((\text{IMT}_1 \leq 22 \text{ hours})\) and \((\text{IMT}_2 \leq 21.75 \text{ hours})\), no daughter cell dies.
2. Else if \((22.75 \leq \text{IMT}_1 \leq 26.5 \text{ hours})\), no daughter cell dies.
3. Else if \((\text{IMT}_2 \leq 17.75 \text{ hours})\), no daughter cell dies.
4. Else if \((23.75 \leq \text{IMT}_2 \leq 26 \text{ hours})\), 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 statistical evidence to say the same in the Second Generation. Surprisingly, the proportion of Deaths in the Second Generation for 1µM is lower than that for Control and 10µM. For Stressed versus Resistant and Full Clonal Potential (FCP) versus Partial Clonal Potential (PCP), there is a trend from Control to 10µM 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 10µM 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, IMT₁, IMT₂ and IMT₃ 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 5 shows the ECDF of DIV for Control, 1µM and 10µM lineages. It shows a big discrepancy among the distributions of Control, 1µM and 10µM starting at 10 hours. 1µM is delayed and 10µM is greatly delayed when compared to Control. The Kolmogorov-Smirnov test finds statistical evidence against the null hypothesis for both Control versus 1µM (p-value < 10^{-3}) and Control versus 10µM (p-value < 10^{-5}). Just to illustrate this difference, 91% of the Control group’s DIV times are less than to 22 hours, but only 73% of the 1µM group’s DIV and 50% of the 10µM group’s DIV are below that threshold. The maximum difference between Control and 1µM ECDFs is 35%, which is at 19.25 hours; and between Control and 10µM ECDFs is 41%, which occurs at 22 hours.

Figure 6 shows the ECDF of IMT_1 for Control, 1µM and 10µM lineages. We observe that 1µM is slightly delayed with respect to Control, and 10µM is slightly delayed with respect to 1µM. The Kolmogorov-Smirnov test finds statistical evidence showing that the distribution of IMT_1 is different in Control and 10µM (p-value < 10^{-5}). 57% of Control IMT_1 are lower than 20.5 hours, but only 44% of 1µM IMT_1 and 18% of 10µM IMT_1 are below that threshold. The maximum difference between Control and 10µM is 39%, which is located at 20.5 hours.

Figure 7 shows the ECDFs of IMT_2. We observe that 1µM is very similar to Control (it almost follows the Control distribution) and 10µM is delayed. The Kolmogorov-Smirnov test finds statistical evidence showing that the distribution of IMT_2 is different in Control group and 10µM group (p-value < 10^{-5}). There is also statistical evidence showing that the distribution of 1µM and 10µM are different (p-value < 10^{-3}). 60% of Controls and 58% of 1µM IMT_2 are less than 19.5 hours, but only 27% of 10µM IMT_2 are less than this threshold. The maximum difference between Controls and 10µM is 33%,
Table 4. Cytotoxic effect in the First and Second Generations for $1\mu M$ and $10\mu M$. Those with statistically different proportions with respect to Control are shown in **boldface**. The values in parentheses are the corresponding p-values.

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

Table 5. The values of proportions from Table 4

<table>
<thead>
<tr>
<th>Proportion</th>
<th>Gen.</th>
<th>Control</th>
<th>$1\mu M$</th>
<th>$10\mu M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death/Non-Death</td>
<td>1st</td>
<td>0.0444</td>
<td>0.1270</td>
<td>0.1184</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>0.1328</td>
<td>0.1111</td>
<td>0.1935</td>
</tr>
<tr>
<td>Stressed/Resistant</td>
<td>1st</td>
<td>0.1966</td>
<td>0.2500</td>
<td>0.3947</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>0.0912</td>
<td>0.1692</td>
<td>0.3276</td>
</tr>
<tr>
<td>FCP/PCP</td>
<td>1st</td>
<td>3.8261</td>
<td>3.2500</td>
<td>0.5714</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>15.0769</td>
<td>10.0000</td>
<td>1.8333</td>
</tr>
</tbody>
</table>

Figure 5. Comparison of the ECDFs of the Progenitor division time
which occurs at 19.5 hours; and between 1µM and 10µM is 32%, which is at 17.5 hours.

Figure 8 shows the ECDFs of IMT\textsubscript{3}. We observe the same trends as in Figure 7: the 1µM group is very close to the Control group and the 10µM group is closer to the Control group than before. The Kolmogorov-Smirnov test finds statistical evidence showing that the distribution of IMT\textsubscript{3} is different in Control and 10µM (p-value \textless 10\textsuperscript{-5}). There is also evidence showing that the distribution of 1µM and 10µM are different (p-value \textless 10\textsuperscript{-4}). 59% of 1µM IMT\textsubscript{3} and 57% of Control IMT\textsubscript{3} are lower than 19.5 hours, but only 29% of 10µM IMT\textsubscript{3} are lower than this threshold. The maximum difference between Control and 10µM is 28%, which is at 19.5 hours; and between 1µM and 10µM 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, 1µM or 10µM.

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 J48\textsuperscript{4} 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\textsubscript{1,1} = Arrested), the lineage is Stressed.
2. Else if (IMT\textsubscript{1,2} \geq 39.5 hours), the lineage is Stressed.
3. Else if (Status\textsubscript{1,2} = Death), the lineage is Stressed.
4. Else if (Status\textsubscript{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\textsubscript{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_{1,2} 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.
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)

<table>
<thead>
<tr>
<th>Gen.</th>
<th>Method</th>
<th>Overall</th>
<th>Stressed</th>
<th>Resistant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Progenitor</td>
<td>JRIP</td>
<td>76.0%</td>
<td>36.8%</td>
<td>90.0%</td>
</tr>
<tr>
<td></td>
<td>J48</td>
<td>65.9%</td>
<td>0.0%</td>
<td>89.4%</td>
</tr>
<tr>
<td>First</td>
<td>JRIP</td>
<td>91.2%</td>
<td>75.4%</td>
<td>96.9%</td>
</tr>
<tr>
<td></td>
<td>J48</td>
<td>92.2%</td>
<td>75.4%</td>
<td>98.1%</td>
</tr>
<tr>
<td>Second</td>
<td>JRIP</td>
<td>94.9%</td>
<td>89.5%</td>
<td>96.9%</td>
</tr>
<tr>
<td></td>
<td>J48</td>
<td>95.8%</td>
<td>89.5%</td>
<td>98.1%</td>
</tr>
</tbody>
</table>

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

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
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

<table>
<thead>
<tr>
<th>Drug</th>
<th>Status_{_3,5}</th>
<th>Status_{_1,2}</th>
<th>Control</th>
<th>Perturbed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Refused</td>
<td>Divided</td>
<td></td>
<td>75.0%</td>
<td>25.0%</td>
</tr>
<tr>
<td>Arrested</td>
<td>Divided</td>
<td></td>
<td>34.6%</td>
<td>65.4%</td>
</tr>
<tr>
<td>Lost</td>
<td>Lost</td>
<td></td>
<td>81.3%</td>
<td>18.8%</td>
</tr>
<tr>
<td>Lost</td>
<td>Divided</td>
<td></td>
<td>47.6%</td>
<td>52.4%</td>
</tr>
<tr>
<td>Never–Born</td>
<td>Unresolved</td>
<td></td>
<td>75.0%</td>
<td>25.0%</td>
</tr>
<tr>
<td>Never–Born</td>
<td>Arrested</td>
<td></td>
<td>87.9%</td>
<td>12.1%</td>
</tr>
<tr>
<td>Never–Born</td>
<td>Death</td>
<td></td>
<td>29.2%</td>
<td>70.8%</td>
</tr>
<tr>
<td>Never–Born</td>
<td>Divided</td>
<td></td>
<td>63.5%</td>
<td>36.5%</td>
</tr>
<tr>
<td>Death</td>
<td>Divided</td>
<td></td>
<td>59.1%</td>
<td>40.9%</td>
</tr>
<tr>
<td>Divided</td>
<td>Divided</td>
<td></td>
<td>80.4%</td>
<td>19.6%</td>
</tr>
</tbody>
</table>

Table 8. Extract of the probability distribution of IMT_{\_1,2} given Drug and Status_{\_2,4} for the network in Figure 11

<table>
<thead>
<tr>
<th>IMT_{_1,2}</th>
<th>Drug</th>
<th>Status_{_2,4}</th>
<th>(0, 19.1]</th>
<th>(19.1, \infty)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>Arrested</td>
<td>70.0%</td>
<td>30.0%</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>Death</td>
<td>10.0%</td>
<td>90.0%</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>Divided</td>
<td>32.9%</td>
<td>67.1%</td>
</tr>
<tr>
<td></td>
<td>Perturbed</td>
<td>Arrested</td>
<td>16.7%</td>
<td>83.3%</td>
</tr>
<tr>
<td></td>
<td>Perturbed</td>
<td>Death</td>
<td>5.0%</td>
<td>95.0%</td>
</tr>
<tr>
<td></td>
<td>Perturbed</td>
<td>Divided</td>
<td>7.8%</td>
<td>92.2%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>DIV</th>
<th>Drug</th>
<th>DIV_D</th>
<th>(0,9.4]</th>
<th>(9.4,20.9]</th>
<th>(20.9,\infty)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>(0,59.4]</td>
<td>29.8%</td>
<td>60.0%</td>
<td>10.2%</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>(59.4,\infty)</td>
<td>3.2%</td>
<td>67.7%</td>
<td>29.0%</td>
</tr>
<tr>
<td></td>
<td>Perturbed</td>
<td>(0,59.4]</td>
<td>50.4%</td>
<td>16.2%</td>
<td>33.3%</td>
</tr>
<tr>
<td></td>
<td>Perturbed</td>
<td>(59.4,\infty)</td>
<td>11.9%</td>
<td>15.3%</td>
<td>72.9%</td>
</tr>
</tbody>
</table>
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 >= 21 hours) and (DIV_D >= 32.25 microns), the lineage is Perturbed.
2. Else if (DIV <= 10 hours) and (IMT_1_2 >= 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, 1µM and 10µM groups, with Control cells having the lowest displacements and 10µM cells having the greatest displacements. For example, 77% Controls, 65% 1µM and 44% 10µM are lower than 37 microns. The maximum difference between Control and 1µM is 26%, which is located at 27 microns. The maximum difference between Control and 10µM is 33%, which is located at 37 microns. The differences between Control and 1µM and between Control and 10µM 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, 10µM 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 1µM cells are delayed more than 10µM 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 1µM 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 1µM

<table>
<thead>
<tr>
<th>DIV_D</th>
<th>Drug</th>
<th>(0,59.4]</th>
<th>(59.4,inf)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>91.3%</td>
<td>8.7%</td>
<td></td>
</tr>
<tr>
<td>Perturbed</td>
<td>66.9%</td>
<td>33.1%</td>
<td></td>
</tr>
</tbody>
</table>
and 10µM 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% 1µM and 44% 10µM are lower than 51 microns. The maximum difference between Control and 1µM is 20% and it is located at 55 microns. The maximum difference between Control and 10µM 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.

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>number of times selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIV</td>
<td>250</td>
</tr>
<tr>
<td>IMT1_1</td>
<td>5</td>
</tr>
<tr>
<td>IMT1_2</td>
<td>250</td>
</tr>
<tr>
<td>DIV_D</td>
<td>250</td>
</tr>
<tr>
<td>IMD2_2</td>
<td>250</td>
</tr>
<tr>
<td>IMD3_8</td>
<td>250</td>
</tr>
<tr>
<td>Status1_2</td>
<td>250</td>
</tr>
<tr>
<td>Status2_2</td>
<td>249</td>
</tr>
<tr>
<td>Status2_4</td>
<td>243</td>
</tr>
<tr>
<td>Status3_4</td>
<td>1</td>
</tr>
<tr>
<td>Status3_5</td>
<td>250</td>
</tr>
<tr>
<td>Status3_8</td>
<td>250</td>
</tr>
</tbody>
</table>

Figure 16 and Figure 17 show this difference for Progenitor and First Generation cells. We observe that 10µM 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 1µM moves slightly faster than Controls and 10µM moves faster than both. This time, both differences are statistically significant with a Kolmogorov-Smirnov test when compared to Control (p-values < 3∙10^{-4}). For example, 73% Controls, 54% 1µM and 49% 10µM are lower than 2.7 micron/hour. The maximum difference between Control and 1µM is 21% and it is located at 2.8 micron/hour. The maximum difference between Control and 10µM 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 10µM, where statistically significant differences between Control and 10µM lineages are found in almost all tests. On the other hand, the effect of the drug in 1µM has only a statistical significant impact in the *Death to Non-Death* proportion and in the ECDFs of DIV and IMT. *Death Non-Death* proportion in the Second Generation is lower for 1µM than both Control and 10µM. 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., 1µM has still some cytotoxic effect, but it is almost lost in the Second Generation, as we see in Figure 7 where 1µM IMT distribution is very close to IMT Control distribution (although its IMD is still affected as shown in Figure 14). The greatest variation between Control and 10µM is found in the proportion of Full Clonal Potential to Partial Clonal Potential. 10µM 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 1µM and 10µM 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 1µM cells, which in turn move less than 10µM cells. This does not happen in the Second Generation, where 1µM and 10µM distributions are interlaced, and over some ranges the 10µM distribution is closer to the Control distribution (see Figure 14).

When comparing IMD and IMT distributions, we observe that in the Second Generation, the 1µM IMT ECDF is very close to Control distribution, but the same does not happen with the 1µM 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, 10µM is the fastest population with the exception of Second Generation (Figure 18) where 1µM is the fastest one. This is related to what we already noted above (see Figure 14) where 1µM and 10µM IMD distributions are similar, but we learn from Figure 18 that 1µM cells are moving faster in a shorter IMT. Also surprising is that the behaviour of 1µM departs from that of the Control group in Second Generation. The behaviour of both 1µM and 10µM 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


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. 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.

2 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; C4.5 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 interdisciplinary 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.