JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2006) **3**, 515–526
doi:10.1098/rsif.2005.0109
Published online 19 January 2006

Integrated mechanistic and data-driven modelling for multivariate analysis of signalling pathways

Fei Hua†, Sampsa Hautaniemi*,†, Rayka Yokoo and Douglas A. Lauffenburger

Biological Engineering Division, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Mathematical models of highly interconnected and multivariate signalling networks provide useful tools to understand these complex systems. However, effective approaches to extracting multivariate regulation information from these models are still lacking. In this study, we propose a data-driven modelling framework to analyse large-scale multivariate datasets generated from mathematical models. We used an ordinary differential equation based model for the Fas apoptotic pathway as an example. The first step in our approach was to cluster simulation outputs generated from models with varied protein initial concentrations. Subsequently, decision tree analysis was applied, in which we used protein concentrations to predict the simulation outcomes. Our results suggest that no single subset of proteins can determine the pathway behaviour. Instead, different subsets of proteins with different concentrations ranges can be important. We also used the resulting decision tree to identify the minimal number of perturbations needed to change pathway behaviours. In conclusion, our framework provides a novel approach to understand the multivariate dependencies among molecules in complex networks, and can potentially be used to identify combinatorial targets for therapeutic interventions.

Keywords: apoptosis; machine learning; mechanistic modelling; signalling pathways; systems biology

1. INTRODUCTION

Despite decades of efforts, therapeutical approaches for complex diseases, such as cancers and autoimmune diseases, remain less effective broadly and more susceptible to detrimental side effects than desired. One of the key reasons hindering the discovery of effective treatments is the daunting complexity of the molecular networks governing cell functional behaviour. Mathematical models are becoming proposed as a powerful new tool to help understand such complexity (Ma'ayan *et al.* 2005). Among a wide array of prospective modelling approaches, knowledge-driven mechanistic modelling using differential equations are in most popular use. In recent years, mechanistic models have provided insights concerning operation of a variety of signalling pathways, such as those featuring NF-κB (Hoffmann *et al.* 2002), the EGF receptor (Kholodenko *et al.* 1999; Schoeberl *et al.* 2002; Wiley *et al.* 2003) and apoptotic caspases (Bentele *et al.* 2004; Hua *et al.* 2005). However, these models, generally formulated as differential equations, are most easily used to study limited facets of the network of interest, especially when attempting to produce conceptual insights concerning how multiple variables work together to yield overall system behaviour. A major reason for the inherent difficulty in going from mechanistic models to network 'logic' is that these models often have tens to hundreds of differential equations and become too large to yield to analytical examination. Commonly used analysis techniques such as sensitivity analysis are typically pursued in univariate mode, where the value of only a single parameter is varied at once. Conclusions drawn from such analyses are hence limited to a particular set of parameter values, to which the model has been fit.

In this study, we propose a framework to facilitate the understanding of the signalling pathway regulations with variations in multiple parameter values, specifically non-zero initial conditions (subsequently referred to as only 'initial conditions') representing molecular component levels. Since signalling molecules are highly interconnected and signal transduction kinetics are nonlinear, the regulation of signalling pathways is inherently multivariate. Moreover, diverse cell types, as well as cells of the same type from different individuals, can differ in expression levels of multiple molecules. Therefore, it is important to study how

*Author for correspondence (sampsa@mit.edu).
†The first two authors contributed equally to this work.

The electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2005.0109 or via http://www.journals.royalsoc.ac.uk.

Received 14 November 2005
Accepted 16 December 2005

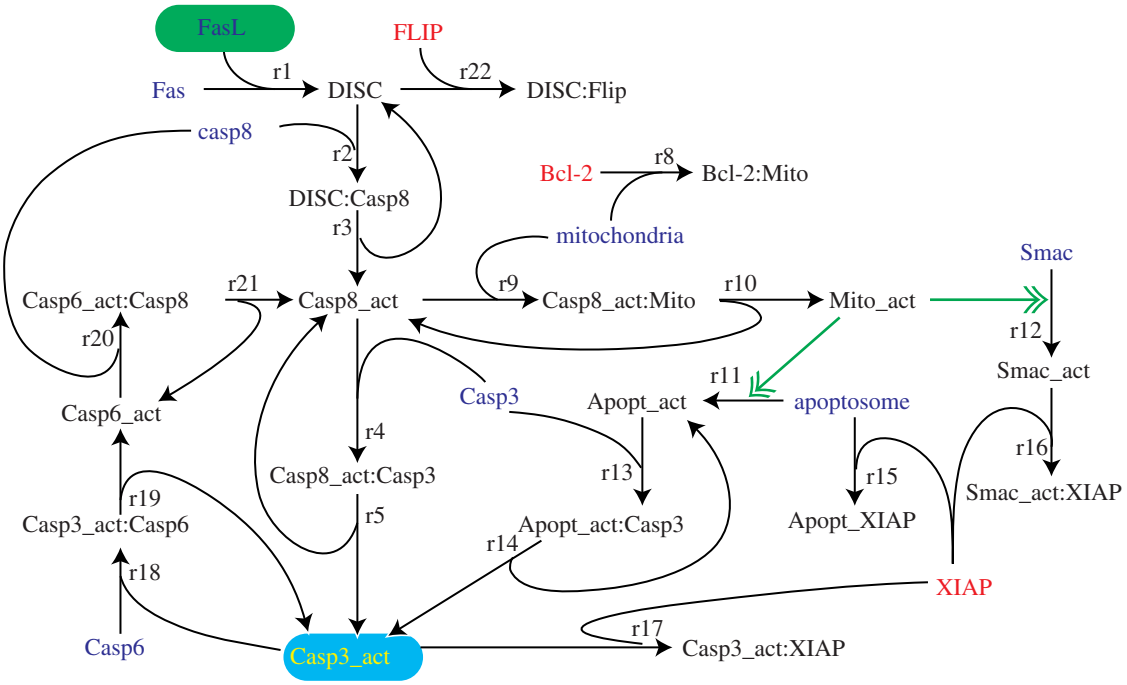

Figure 1. Schematic representation of the Fas pathway used in the mechanistic model. Molecules in blue are directly involved in Fas signalling; molecules in red are inhibitors of the pathway; intermediate reaction complexes are in black. The triggering molecule, FasL, is shaded in green. The output molecule is shaded in blue. Two green coloured reaction arrows represent transport type reactions. Numbers on the arrow indicate the corresponding reactions in electronic supplementary material.

multiple components integrate to regulate the signalling pathways in order to appreciate different responses to the same stimuli among different cells. Our proposed framework meets this challenge by applying a data-driven modelling approach to analyse the multivariate dataset generated from a mechanistic model.

Mechanistic models are knowledge-driven in nature, requiring considerable amount of knowledge about the biological process to be specified, including the pathway connectivity, initial conditions and reaction rate constants. A central advantage of this approach is to offer quantitative and temporal (and even spatial in cases) information of the signalling network *in silico*. These models have the potential to allow studies of multivariate regulations by running a large number of simulations with combinations of varied initial conditions. However, we are not aware of any effort to date for extracting useful regulation information from a 'library' of multivariate simulations.

In contrast, data-driven models are derived directly from measurements (or calculations) describing input and/or output characteristics of a system. Such approaches enable extraction of general and abstract rules governing biological processes from a large amount of data. The data used in data-driven models are mostly high-throughput experimental data, such as microarray data and mass spectrometry data (Adam *et al.* 2002; Segal *et al.* 2003; Basso *et al.* 2005). In the present work, we apply a data-driven modelling approach to the simulated data from a mathematical model.

Our proposed data-driven modelling approach consists of two steps: clustering of mechanistic model simulation outputs and subsequent classification of the clustered outputs based on initial conditions. The goal of the clustering step is to group simulation outputs into discrete subsets (i.e. clusters) and these clusters are then used as outcomes in the classification step. The clustering step is essential since classification requires discrete outputs. Moreover, clustering reduces the dimensionality of the output space, which facilitates extraction of generalized relationships between inputs and outputs. In the classification step, we use a decision tree algorithm to identify the molecules that together are able to predict different outcomes (Breiman *et al.* 1984). Initial conditions for each simulation are the inputs of the decision tree algorithm and the clusters based on the simulation outputs are the outcomes (the leaf nodes of the tree). The decision tree algorithm repeatedly splits the dataset based on selected input values—in this case, molecule concentrations—to maximize outcome purity of resulting data subsets. Consequently, the algorithm generates a tree graph, which illustrates propositional (Boolean) rules leading to a certain cluster. The rules are composed of a set of initial condition ranges. According to the tree graph, the set of key molecules and their concentration ranges that lead to certain pathway behaviour can be clearly read out by following a path from the root of the tree to a leaf.

Decision tree analysis has been applied to many different areas of biomedical decision making processes (Adam *et al.* 2002; Podgorelec *et al.* 2002; Hautaniemi *et al.* 2005). In this study, we take a step further by using the decision tree model to predict manipulations needed to change cellular responses to the activation of the pathway. Since molecules and their concentration ranges are given for each path of the decision tree, we are able to identify the molecules that have different values on two paths leading to two different behaviours.

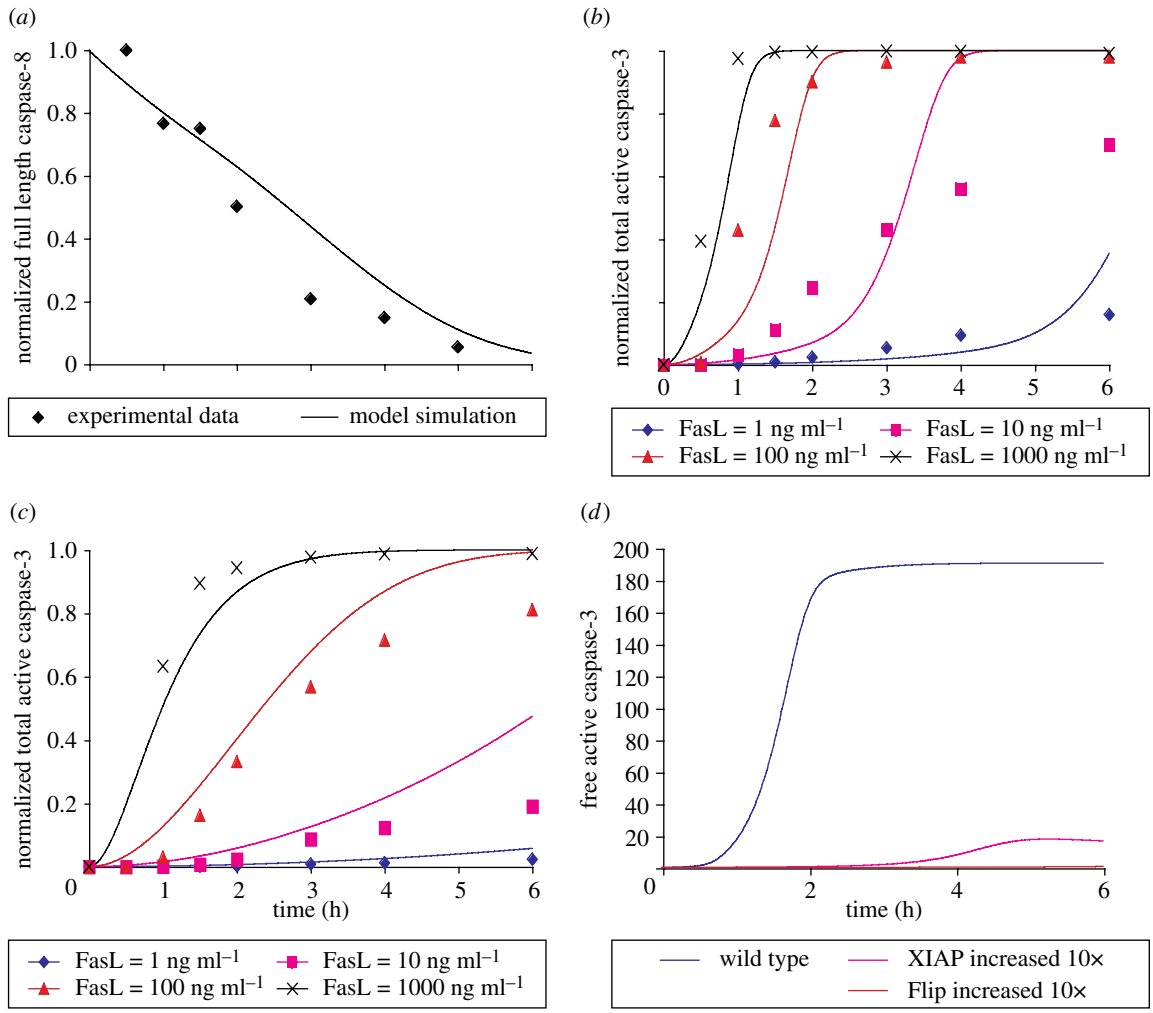

Figure 2. Model fitting with experimental data. Separated symbols are experimental data and smoothed curves are model simulations. (*a*) Model fitting for full-length caspase-8 decrease in the presence of 100 ng ml^{-1} FasL in control Jurkat cells. (*b*, *c*) Model fitting for total active caspase-3 generation time-course in the presence of 1, 10, 100 and 1000 ng ml^{-1} FasL in control Jurkat cells (*b*) or Bcl-2 overexpressed Jurkat cells (*c*). (*d*) Model simulation shows slow down of caspase-3 activation with 10 times increased XIAP or Flip in the presence of 100 ng ml^{-1} FasL.

These are the predicted molecules that need to be perturbed to switch a cell from one path to the other. This approach offers the potential for identifying combination of perturbations required to change cellular responses.

We study here the Fas apoptotic pathway as an example system for our integrated analysis. Fas is a member of death receptor family, which transduces the extracellular cue Fas ligand (FasL) in governance of programmed cell death. Fas-induced apoptosis plays an important role in many cellular processes, dysregulation of which can lead to various diseases such as cancer and autoimmune diseases (Landowski *et al.* 1997; Krammer 2000). As a tightly controlled cell suicide process, the signalling molecules downstream of Fas and FasL binding have been well characterized in isolated pieces (Nagata 1999; Igney & Krammer 2002; Lavrik *et al.* 2005) and several mechanistic models for this pathway have also been developed (Fussenegger *et al.* 2000; Bentele *et al.* 2004; Hua *et al.* 2005). Nonetheless, it remains unclear how this network is regulated under different combinations of relevant protein expression levels. By applying our integrated data-driven modelling with mechanistic modelling framework to this pathway, we observe that whether varying the concentration of a single molecule affects the pathway output is strongly dependent on the concentrations of other molecules in the pathway. Similar network outputs or cellular responses can result from different subsets of components exhibiting diverse combinations of expression levels. Consequently, manipulations of different protein(s) might be required to alter similar responses with different underlying mechanisms.

2. RESULTS

2.1. Generation of a simple mechanistic model for the Fas signalling network in Jurkat cells

We created a simplified mechanistic model describing the dynamic signal propagation for the Fas pathway derived from a previously published model (Hua *et al.* 2005). To keep the number of parameters in the model small, the previous model was simplified by aggregating several molecules in a linear pathway into a single surrogate molecule at several locations in the network while the basic structure of the network, including both

the direct and mitochondria-involved pathways, was maintained. Figure 1 shows the network structure described by the simplified model. The pathway is triggered by adding FasL and ends at the activation of caspase-3. Since the active caspase-3 can be sequestered by free X-linked inhibitor of apoptosis protein (XIAP) and become inactivated, the kinetics of free (i.e. XIAP unbound) cleaved caspase-3 production was used as the model output. An extensive experimental dataset was measured for parameter fitting, including a time-course for caspase-8 cleavage in control Jurkat cells stimulated with 100 ng ml^{-1} FasL, and time-courses of total active caspase-3 production in control and Bcl-2 overexpressed Jurkat cells stimulated with 1, 10, 100 and 1000 ng ml^{-1} FasL. Figure 2*a–c* shows model simulation results after parameter fitting. In addition to fitting the model with experimental data, the model also predicted delayed caspase-3 activation when the concentration of Flip or XIAP is increased (figure 2*d*), which are consistent with literature data (Irmler *et al.* 1997; Bratton *et al.* 2002). All the biochemical reactions, reaction rate constants and non-zero initial conditions (concentrations of the molecules before adding FasL excluding transient complexes and molecules) used in the model are listed in electronic supplementary material.

2.2. Generation of multivariate simulation dataset

In order to understand the regulation of the Fas pathway with different concentrations of the molecules in the pathway, it is desirable to run simulations with combinations of varied initial conditions for multiple molecules. Depending on how many potential values are given to each initial condition, covering all the possible combinations leads to an exponential increase of the number of simulations. To overcome the challenge of computational complexity, we performed Monte Carlo simulation with randomly chosen values for 9 out of 11 initial conditions; the excluded molecules were FasL and caspase-3. The initial condition for FasL was kept the same throughout the simulations because it is an external signal added to the cell, therefore, the regulation of the Fas pathway by a cell itself does not include controlling FasL concentration. The initial condition of caspase-3 was kept constant because it is the proenzyme of the output product, therefore, changing it obviously will have a dramatic effect on the model output.

All the varying initial conditions exhibit a log-uniform distribution within a range of 10 fold higher or lower than the baseline values, which are the values in the model fitted to Jurkat cells. Taking the log-transformation ensures that the number of possible states with values less than the baseline value is the same as the number of states with values greater than the baseline. A total of one million simulations were run using MATLAB. Each simulation represents a single *in silico* cell with an independent set of initial conditions. All initial conditions are expressed as fold changes with respect to the baseline values.

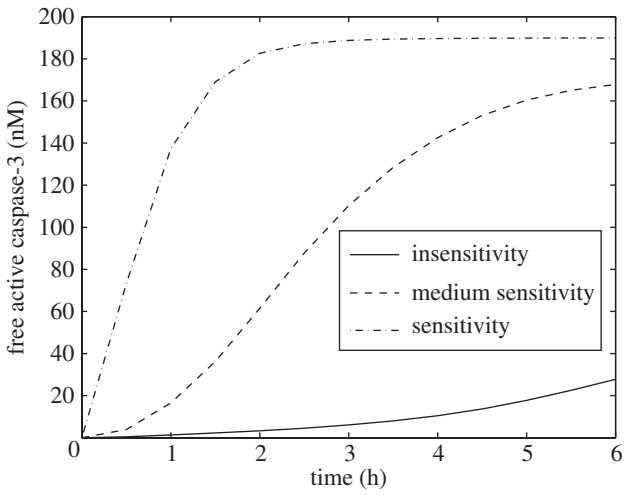

Figure 3. The centroids of free cleaved caspase-3 production curve for each cluster.

2.3. Data clustering

The model outputs from one million simulations varied widely from very fast with free caspase-3 concentration reaching the highest concentration in 30 min to almost no free caspase-3 generation during the length of simulation. In order to explore this large dataset, we first transformed the dataset into discrete clusters using the *k*-means clustering algorithm (Duda *et al.* 2001). The clustering was based on the model output, i.e. time-courses of free cleaved caspase-3 production. Since free cleaved caspase-3 cleaves many important cellular substrates leading to morphological changes that are typically associated with apoptosis, we used this model output to indicate the cellular response to Fas-induced apoptosis. A fast increase in free cleaved caspase-3 concentration and/or a large steady state concentration implies that the cell is sensitive to Fas-induced apoptosis and vice versa. We chose three clusters for the clustering algorithm to represent insensitive, medium sensitive and sensitive cellular responses to Fas-induced apoptosis. Figure 3 shows the resulting centroids of the free cleaved caspase-3 production curve for each cluster.

2.4. Decision tree analysis of multivariate regulations of the Fas pathway

To extract and visualize the complex relationships among all the molecules in the Fas pathway, we applied the decision tree analysis to the clustered simulation dataset as described in §4. The resulting decision tree graph is illustrated in figure 4. The top node splits the whole dataset depending on whether the initial condition for XIAP in each cell is less than 3.1 fold change from the baseline value. At each internal node, the left edge is taken if the initial condition of the denoted molecule is within the given range; otherwise, the right edge is taken. Each leaf node is the predicted outcome according to the decision tree model, which is one of the three clusters of the cellular sensitivity to FasL stimulation. A path from the top node to a leaf node comprises a set of rules, which are the value ranges

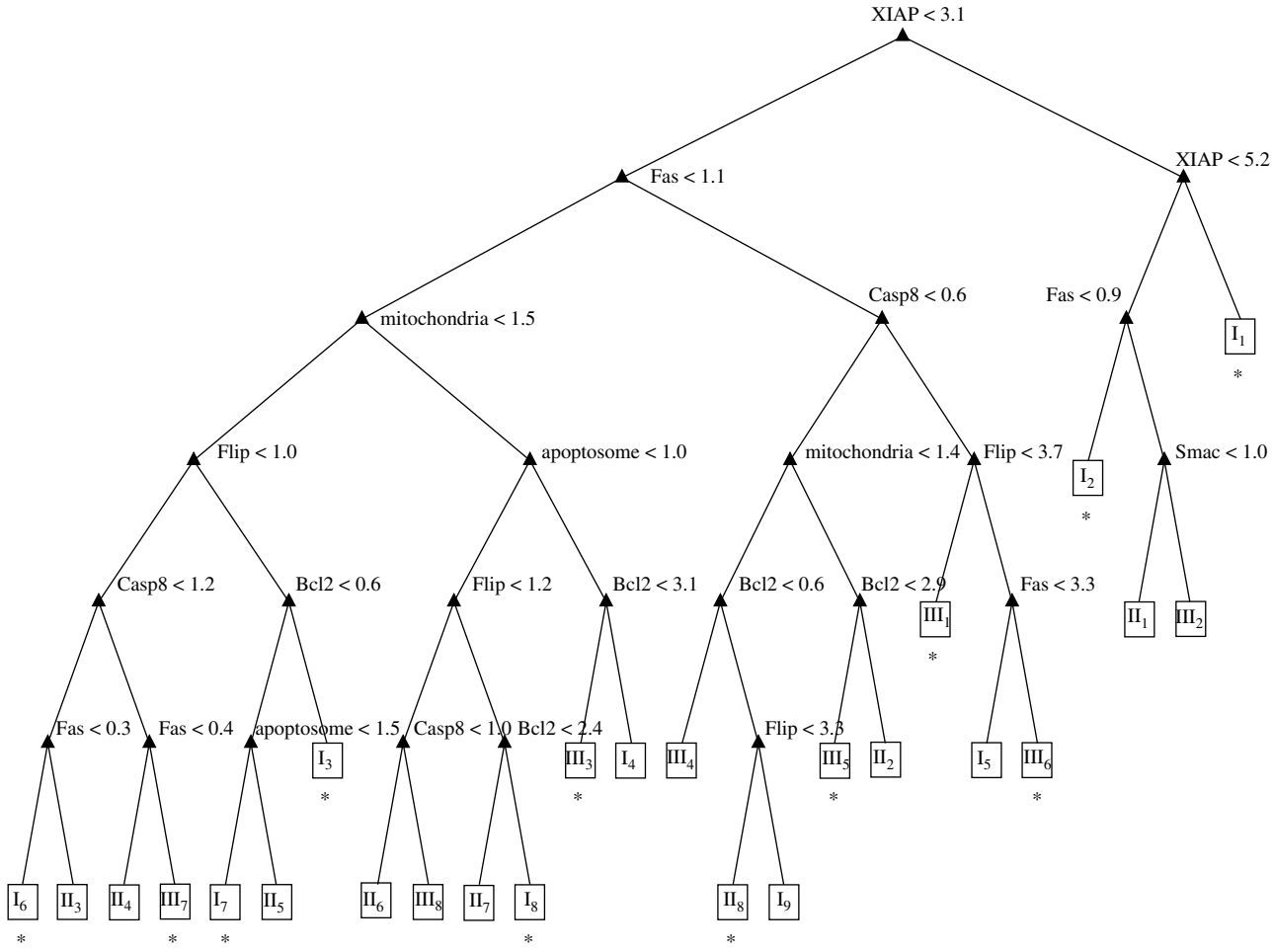

Figure 4. Decision tree model based on the Fas mechanistic model simulation data. The internal nodes have a triangle shape and the leaf nodes have a square shape. Roman numerals I, II and III for the leaf nodes correspond to the insensitive-response, medium sensitive-response and sensitive-response clusters, respectively. Subscript numbers for the leaf nodes, which are ordered from top to bottom, are used to identify each leaf node. The leaf nodes with over 66% prediction accuracy (in the training set) are marked with asterisks.

of multiple initial conditions. The components shown on the path are defined as important (key) molecules for the outcome of the path because their concentrations are pertinent. For example, for the fourth leaf from right (labelled with I_2), the decision tree predicts that if a cell has XIAP level between 3.1 and 5.2 times its baseline value and Fas level below 0.9 times its baseline value, the cell belongs to the insensitive-response cluster. Therefore, XIAP and Fas are key components in determining the insensitive-response behaviour, while the concentrations of other molecules do not matter significantly in this case.

From the emerging decision tree model, we can see that similar behaviour (i.e. the same cluster) can be caused by different combinations of key molecules and different initial concentration ranges for even the same molecule in different sets. The decision tree model also suggests that the importance of one molecule depends on the relative concentrations of other molecules. For instance, Smac is only important when XIAP is at an intermediate high level (between 3.1 and 5.2 times baseline level) and Fas is also high. When the XIAP level is very high (greater than 5.2 times of the baseline value), the cell is always insensitive to Fas killing because XIAP level greatly exceeds the highest possible level of Smac so that there are always plenty of Smac-unbound XIAP to sequester newly generated cleaved caspase-3. When the XIAP level is very low (less than 3.1 times the baseline value), Smac is not critical either. This is because the sole function of Smac in this model is to reduce the amount of available XIAP that sequesters caspase-3 and -9, while the XIAP level is so low that it is not sufficient to sequester caspases anyway. Another example is Bcl-2. In the decision tree, Bcl-2 appears in many paths, but not all. This suggests that the importance of Bcl-2 in regulating cellular response to FasL stimulation is highly dependent on the levels of the other molecules. This gives a potential explanation for the existence of cells that are either sensitive or insensitive to Bcl-2 overexpression (Scaffidi *et al.* 1998).

We compared the decision tree graph to the sensitivity analysis results (figure 5) from the mechanistic model with baseline initial conditions. The sensitivities of the model output to individual initial conditions were calculated in Jacobian (Numerica Technology, LLC, Cambridge, MA), in which the sensitivity to perturbations of each single initial condition is calculated independently. The pathway

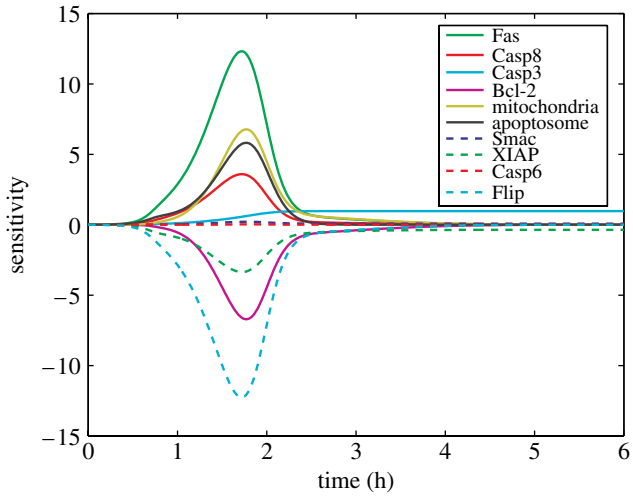

Figure 5. Sensitivity analysis for the Fas mechanistic model with all the baseline initial conditions. Sensitivity of the free cleaved caspase-3 in response to changes of each single initial condition over time is calculated.

output appears to be most sensitive to the changes of initial conditions for Fas and Flip molecules given all the initial conditions are at the baseline value. In contrast, the decision tree algorithm is applied to the dataset with simultaneous variation of concentrations for multiple molecules. It identifies subsets of proteins whose concentrations together are most predictive of the pathway outcome. Fas and Flip, the most sensitive molecules from the sensitivity analysis, are also present on many paths, which suggest that their concentrations are important with many different sets of initial conditions. However, under some conditions (e.g. paths leading to leaf nodes III_1, III_5 and II_8), other molecules such as caspase-8, Bcl-2 or mitochondria are also important. Since our simulation dataset covers a wide range of concentration for each molecule, we were not only able to identify these key molecules, but also specify ranges where the molecule concentrations need to fall into.

2.5. Validation of the decision tree model

To validate the decision tree model (figure 4) we created a second Monte Carlo dataset consisting of one million new simulations. The distances between the output of each simulation in this validation dataset and all three previously determined k-means centroids were calculated using the cityblock distance. The cluster giving the closest distance was assigned to each simulation. We then tested how well the decision tree model can predict the cluster outcome according to the initial conditions for each new simulation. Since there are three possible clusters outcomes, we would expect to achieve 33% average accuracy by randomly choosing a cluster. The decision tree resulted in 71% prediction accuracy for this independent testing dataset, which is more than twice as good as randomly choosing the cluster.

For training and validation Monte Carlo simulations, we set the sampling interval for each initial condition to be $10^{0.025}$. Given the fold-change from the baseline value with a range from 10^{-1} to 10^{1}, each initial condition has 81 potential values. To test whether the sampling interval for each molecule concentration is small enough, we generated a third Monte Carlo dataset with a larger sampling interval of $10^{0.25}$, while keeping the ranges of the initial conditions the same resulting in nine potential values for each molecule. Using this lower resolution dataset of one million simulations as the testing dataset, the decision tree model in figure 4 still resulted in 71% prediction accuracy, which suggests that the original higher resolution dataset is sufficient to capture the essential characteristics of the space of all variable combinations.

2.6. Identification of minimal perturbations to switch the behaviour of the Fas pathway

In addition to showing the important multivariate relationships among the molecules in the Fas pathway, we used the decision tree model to make predictions about manipulations needed to change the cellular sensitivity to FasL stimulation. By following a path in the decision tree, the sets of important molecules that lead to certain cellular response cluster and their concentration ranges can be clearly read out. Therefore, we were able to identify the molecules with different concentrations between two paths. These are the molecules whose concentrations need to be modified to change cellular responses. The total number of these molecules is defined as COST and the detailed COST calculation is given in §4. The calculation of COST is direction-dependent, i.e. the COST from path A to path B may be different from the COST from path B to path A. We computed COST values between all the combinations of any two paths that do not belong to the same cluster. Table 1 gives the COST matrix for changing from the insensitive-response cluster to the sensitive-response cluster. Five additional COST matrix tables for other combinations are given in electronic supplementary material.

Using the decision tree and the COST matrix, we can suggest how to switch a cell response to FasL stimulation with minimal number of perturbations. First, we use the decision tree to identify which path the cell follows. Second, we use the COST matrix to identify which path to switch to in order to obtain the minimal COST. For example, assume we would like to switch a cell from insensitive to FasL stimulation to sensitive. According to the decision tree, we need to first find out the fold difference of XIAP concentration in this cell relative to Jurkat cells (the fitted baseline model). Depending on the XIAP concentration in the cell, we either know the cell is in leaf node I_1, or we need to measure relative Fas concentration, and so forth. Assuming we eventually find out that the cell follows the path leading to leaf I_5, we use the COST matrix in table 1 to find out that the minimal COST value in the column corresponding to I_5 is one by switching to either leaf node III_1 or III_6. Accordingly, we predict that only one molecule needs to be perturbed which can either be decreasing Flip level or increasing Fas level. As a second example, if a cell also has insensitive response and belongs to leaf I_6, the best way to switch it into

Table 1. COST matrix for transition from insensitive-response cluster (I) to sensitive-response cluster (III).

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9
III_1	2.18	2.67	1.92	1.67	1	1.77	1.92	1.94	1.95
III_2	2.01	1.52	2.5	2.5	1.52	2.51	2.49	2.5	1.52
III_3	2.98	2.49	2	1	2.5	1.87	1.92	1.88	2.99
III_4	3.31	3.79	2.65	3.61	2.18	1.96	1.65	3.6	1
III_5	3.07	3.55	3.07	2.61	1.95	2.61	2.6	2.48	1.46
III_6	2.98	3.22	1.99	2.22	1	2.77	1.99	1.96	1.59
III_7	3.39	3.23	2.16	2.67	2.76	2	2.17	3.17	3
III_8	3.64	3.15	3.01	2.03	3.3	2.47	2.65	1.54	4.51

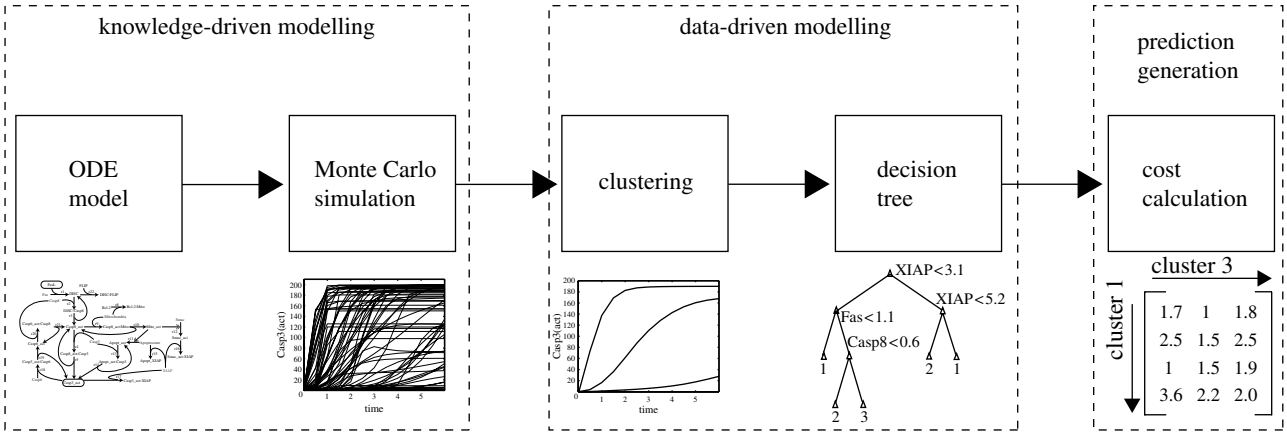

Figure 6. Summary of the framework to understand the multivariate regulation of signalling pathways. The framework starts with the construction of a mechanistic model, which is used to generate a Monte Carlo dataset. The model simulations are then clustered based on the model outputs. These clusters are the outcomes in the decision tree model, while the predictors are initial conditions for each simulation. Finally, according to the decision tree model, the number of perturbations needed to change from one path to another is calculated.

a sensitive cell is to change to leaf III_1 with COST of 1.77 by increasing Fas and possibly increasing caspase-8 as well. Since the caspase-8 concentration range is partially overlapping between two paths (less than 1.2 for leaf node I_6 and ≥ 0.6 for leaf node III_1), there is a possibility that caspase-8 concentration needs to be increased to be greater than 0.6. This possibility is estimated as described in §4.

According to the COST matrix, a minimal COST greater than one is not uncommon for switching between insensitive-response cluster (I) and sensitive-response cluster (III). This suggests that several molecules need to be perturbed simultaneously to obtain a sufficient large change in cellular responses, while changing one molecule may be only capable of switching between two adjacent clusters (e.g. from the insensitive-response to the medium sensitive-response). Taken together, the decision tree model exhibits a promising capability to identify combinatorial perturbations required to obtain a stronger effect.

3. DISCUSSION

In this study, we have developed a framework to study multivariate regulation of cellular signalling pathways *in silico*, summarized in figure 6. This framework integrates a data-driven modelling approach with a knowledge-driven modelling approach to identify the complex relationships between signalling molecule concentrations and the signalling pathway output dynamics and to generate experimentally testable predictions.

Our approach begins with the construction of a knowledge-driven mechanistic model for the pathway, followed by Monte Carlo simulation with varied initial concentrations of signalling molecules. The simulated data are then grouped into three clusters based on the model output, followed by decision tree analysis. Using the propositional rules emergent from the decision tree model, the perturbations needed to switch from one phenotype to another are predicted, and the minimal number of perturbations needed is also identified.

Employing the Fas apoptotic signalling pathway as our example system here, application of this aggregated modelling approach suggests there is no single or fixed subset of molecules that can determine the behaviour of the Fas signalling pathway. Instead, concentrations of different sets of molecules can be important in determining the cellular responses to FasL stimulation, depending on the concentrations of other molecules involved in the Fas signalling. These dependencies are described by the decision tree model. Consequently, similar responses to FasL stimulation in different cells might be due to combinations of different molecule concentrations: thus, these cells require different perturbations to change their behaviours.

3.1. Generation of multivariate simulation dataset with the mechanistic model

Mechanistic models are becoming more and more common in biomedical studies aiming at identifying therapeutical targets (Schoeberl *et al.* 2002; de Pillis *et al.* 2005). Often, analyses of the pathway regulation—such as sensitivity analysis—study the effect of changing one single initial condition on the model output at a time. This renders the analysis results dependent on the other initial conditions chosen for the simulation, and thus to a particular cell type that the model has fit to. One potential approach to overcome this limitation is to run many simulations with combinations of different initial conditions. A major challenge with this approach is an exponential explosion of the possible combinations of initial conditions when there are tens or even hundreds of initial conditions in the model.

One of our strategies to minimize the problem of combinatorial complexity is to use a simplified model to reduce the number of initial conditions in the model. The model is simplified by using functional modules, such as the mitochondrial module and the apoptosome module, based on the biological knowledge of the pathway. Each module represents several molecules whose functions are separable from those of other molecules. This model reduction facilitates identification of important modules and if the understanding of individual molecules in a module is desired in the future, we can generate an expanded model including details for that particular module or apply separate analysis to a particular module.

Our second strategy to overcome the combinatorial complexity is to use Monte Carlo simulation. The number of potential values for each molecule is another factor contributing to combinatorial complexity. Since the relationship between molecule concentrations and model outputs is nonlinear, we would like to have as many values for each molecule and thus smaller intervals between values as possible to capture the appropriate behaviour. However, even five values for each of the nine varying molecules result in almost two million (5^9) combinations. Therefore, we utilized Monte Carlo simulation to randomly select combinations of initial condition values with 81 potential values for each initial condition. Two independent runs of Monte Carlo simulation, with one million individual simulations for each, resulted in very similar decision trees with equal prediction accuracy (results not shown), which suggests that although one million simulations only covers a very small fraction (about $10^{-10}\%$) of all the possible combinations, the decision tree based on these simulations is both representative and reproducible.

Our ordinary differential equation (ODE) model is a greatly simplified depiction of the Fas-induced apoptosis process. It serves well the purpose of this study: developing a machine-learning approach to analyse large-scale dynamic dataset. For future, more dedicated application efforts, the model for this process could be expanded in depth and scope by disaggregating some of the components and including additional events such as protein degradation and synthesis. It would also be of great interest to use some systematic mathematical approach for the model reduction. However, these represent different topics that are outside of the range of this study.

3.2. Data-driven modelling

Since data-driven models have the capability of processing large amounts of data, they provided a good solution to analyse our Monte Carlo dataset. The data-driven modelling framework we have developed consists of clustering and classification steps.

An essential step in our framework is to cluster simulations according to the model output since the clustering results are used as the outcomes in the subsequent decision tree analysis, which requires limited number of discrete outcomes. We chose the k-means algorithm because it is computationally efficient compared with other clustering algorithms, such as hierarchical clustering. It would be an interesting topic for further research to test how different clustering algorithms may affect the decision tree model. One particular clustering algorithm we would like to test in the future is the fuzzy k-means clustering, which computes the likelihood for a simulation belonging to each cluster instead of strictly assigning each simulation to one cluster. The use of fuzzy k-means algorithm would, however, require significant modifications to the decision tree algorithm, which is beyond the scope of this study.

We used the decision tree analysis in the classification step to link the molecule concentrations to simulation outputs. Among a number of standard machine learning classification algorithms, such as artificial neural networks or support vector machines, decision trees possess several attractive properties for extracting key molecules regulating multivariate signalling networks. These advantages include no need to assume a linear relationship or pre-specified distributions for the data. Moreover, the decision tree analysis automatically recognizes the dependence of one molecule on other molecules (i.e. context dependence). Most importantly, the set of key molecules and their value ranges that lead to certain cluster can be clearly read out from the tree. The tree graph illustrates the multivariate relationships among signalling molecules. Regardless of the positions of the molecules in a path (closer to the top or bottom), all the molecules in one path are equally important and their concentrations together determine the outcome of the path.

Since the decision tree analysis illustrates the key molecules in the pathway, it provides a guide as to which and how many molecules we need to measure to predict a cell's response to the activation of the pathway. For instance, instead of measuring all the molecules on the pathway, according to the decision tree, we should measure XIAP level in the cell first, if it is medium high, we should measure Fas, and so forth. At the same time, one should pay attention to how close the concentration of the specified molecule in a cell is to the splitting value. If the concentration of a molecule in the sample to be classified is very close to the splitting value, both branches are almost equally valid (because

each splitting value is an estimate with some level of uncertainty) but they may lead to different classes. As an instance, the simulation for the baseline model, for which all the initial conditions are at onefold change, belongs to the sensitive-response cluster (III). However, it was misclassified to the insensitive-response cluster (I) by following the path leading to leaf node I_3. The reason for this misclassification is due to the partition at the internal node for Fas (Fas <1.1, or more precisely Fas <1.091). Since in the baseline simulation Fas has the value one, both branches could be followed and taking the right branch leads to the leaf node III_1, which is the correct class.

The decision tree model for the Fas pathway shown in figure 4 has an encouraging prediction accuracy of 71%. The decision tree model was better at predicting cells belonging to the insensitive-response (I) and the sensitive-response (III) clusters than cells belonging to the intermediate (II) one. For the sensitive and insensitive leaves, there are only six cases in which the prediction accuracy is under 66%, whereas for medium sensitive-response cluster (II) all but one leaf have prediction accuracies are below 66%. The overall mean prediction accuracy for the leaves with equal to or greater than 66% accuracy (marked with asterisks in figure 4) is 78%, and for the leaves with less than 66% accuracy is 48%. Thus, the majority of the rules for leaves I and III are fairly accurate, such as leaf III_1 with 92% prediction accuracy and leaf I_3 with 90% prediction accuracy.

The prediction accuracy can probably be improved by using ensemble decision tree approaches such as the random forest algorithm (Breiman 2001*a*). However, these new approaches often lose effective graphic representation of the relationships between the initial conditions and the cellular responses. In general, there is an counteracting relationship between the accuracy and the interpretability of a model: a very high prediction accuracy often requires a more complex model, which is in turn more difficult to interpret without extensive computational effort (Breiman 2001*b*). Accordingly, we chose our decision tree analysis with a decent prediction accuracy and excellent interpretability.

3.3. Predicting manipulations according to the decision tree

Since the decision tree provides a convenient visualization of the key molecules leading to certain behaviours, we can directly compare two paths and identify which components need to be changed and by how much in order for the system behaviour to switch from one path to another path. In this study, we have simply calculated the minimal number of perturbations required to switch the pathway behaviour. For example, according to the decision tree (figure 4) and the COST matrix (see electronic supplementary material), the best way to switch Jurkat cells from the sensitive-response cluster (III_1) to an insensitive-response cluster is to increase Flip level, and therefore, switch Jurkat cells to leaf node I_5. This is consistent with literature which has shown overexpression of Flip protects cells from Fas-induced apoptosis (Irmler *et al.* 1997); and also consistent with the ODE model simulation which has shown that increasing Flip can slow down caspase-3 activation in the model for Jurkat cells (figure 2*d*).

In practice, it is much more complex to define the best way to change the cellular behaviour. Possible considerations include the magnitude of the concentration change, how easy it is to manipulate the molecule(s) and whether changing these molecules affects other signalling pathways leading to undesired side effects. The first two criteria can be potentially incorporated into the COST calculation by multiplying a coefficient to each changing molecule indicating either the magnitude of the change or the difficulty of the manipulation. The third consideration is much more challenging. One approach could be to construct separate mechanistic models for other pathways with cross-talk molecules to the Fas pathway and then apply our developed framework described in figure 6 to obtain a decision tree model for each new pathway. These decision tree models can be used to check whether changing a particular cross-talk molecule will affect cellular responses to these pathways. Another more challenging and interesting future direction is to build multi-dimensional decision trees based on a large mechanistic model including several pathways with cross-talk to each other. In this case, the initial conditions would still be the input for the decision tree but the outcomes would be the cellular responses to each pathway.

3.4. Significance of the study

Knowledge-based mechanistic models and data-driven models are two types of computational approaches that complement each other. By combining them to analyse molecular signalling pathways, we have developed a way to obtain multivariate regulatory information and to predict combinatorial perturbations that modify the pathway behaviour.

Due to the heterogeneity among individuals, the protein expression levels in the same cell type may vary considerably across different individuals. Furthermore, when a drug is administered to a patient, it interacts with many different cell types in the human body. These different cell types often have different amounts of protein expression levels as well. Therefore, understanding regulation of molecular signalling pathways with varied signalling molecule concentrations would help us understand different responses among different cell types, as well as the same cell type among different individuals. This knowledge may consequently facilitate personalized treatment and better drug design with minimized side effects. There has been considerable progress in genomic and proteomic research to rapidly identify differences among cells at the RNA level and the protein level (Rosenwald *et al.* 2002; Gunther *et al.* 2003; Kislinger *et al.* 2005). Investigations have been mainly focused on identifying useful biomarkers to predict the effectiveness of a drug treatment or to classify diseased states. Our study provides a new approach to facilitate the translation from the differences in the basic constituents of cells to the differences in cellular responses based upon the knowledge of the

signalling network. Furthermore, potential benefit of targeting multiple targets for effective treatment of diseases has been recognized increasingly (Roth *et al.* 2004; Frantz 2005; Mencher & Wang 2005). The capability of identifying combinatorial targets for effectively changing cellular responses can make the proposed approach very useful in drug design.

This data-driven strategy to analyse and visualize large-scale datasets is versatile and applicable to a wide range of biomedical applications involving prediction, such as diagnosis, prognosis or predicting cell decisions based on intra- or extracellular conditions. Also, when experimental data about different protein expression levels and cellular responses among different cells become available, our data-driven modelling approach can also be applied to analyse these experimental data.

4. MATERIAL AND METHODS

4.1. Cells and reagents

A human tumour T cell line, Jurkat.E6, was purchased from ATCC (Manassas, VA). Bcl-2 overexpressed Jurkat.E6 cells were generated using retrovirus infection as described previously (Hua *et al.* 2005). Recombinant human superFasL was purchased from Alexis (San Diego, CA). Anti-cleaved caspase-3 mAb (552 597) used for intracellular staining was purchased from BD Pharmingen (San Diego, CA), and Alexa Fluo 647 donkey anti-rabbit IgG (A31573) was purchased from Molecular Probe.

4.2. Intracellular staining of active caspase-3

Resting or stimulated cells were fixed with 4% paraformaldehyde for 10 min at room temperature, followed by permeabilization with 100% MeOH overnight at $-20\,^{\circ}\mathrm{C}$. Cells were washed twice with PBS + 0.1% Tween (PBST) before incubating with anti-cleaved caspase-3 for 1 h. Cells were washed twice again with PBST, followed by incubation with Alexa Fluo 647 anti-rabbit at room temperature for 1 h in the dark. After two more washes with PBST, cells were analysed on a FACS Calibur machine (BD Biosciences, San Jose, CA). The percentage of cells positive for cleaved caspase-3 was normalized using following equation:

$$[(\text{cleaved caspase-3}) - (\text{spontaneous cleavage})\%]/[1 - (\text{spontaneous cleavage})\%],$$

where (spontaneous cleavage)% is the percentage of cells that are positive for cleaved caspase-3 without adding FasL.

4.3. Mechanistic model for the Fas pathway

An ODE based mathematical model was constructed to describe the signal transduction along the Fas pathway. Figure 1 schematically illustrates the pathway topology described in the model. The model starts with FasL binding to Fas and ends at the activation of caspase-3. Since the cleaved caspase-3 can be sequestered by XIAP and become inactive, the free cleaved caspase-3 was used as the model output. In the model, the pathway was simplified by aggregating several molecules on a linear pathway into a single surrogate molecule, while still keeping the basic two-branched pathway structure. The model reduction was subjectively determined based on the biological knowledge of the pathway structure. Specifically, Fas molecule in the model aggregates both Fas and FADD; the mitochondria molecule in the model aggregates the signal from Bid activation through mitochondrial membrane disruption, until cytochrome *c* release; the apoptosomes in the model represent the complexes of released cytochrome *c*, Apaf-1, caspase-9 and ATP. Three negative regulatory molecules of the pathway, Flip, Bcl-2 and XIAP were included in the model. Mass action type of biochemical reactions was used to represent protein–protein interactions and enzymatic reactions, except for two transport reactions as indicated.

The model was created in both Jacobian (Numerica Technology, LLC, Cambridge, MA) for the purpose of parameter estimation and sensitivity analysis, and MATLAB (The Mathworks, Inc., Natick, MA) for the purpose of Monte Carlo simulation. A copy of the original code can be obtained upon request. The simplified ODE model consists of 11 molecules of nonzero initial conditions (concentrations), 18 intermediate complexes with zero initial conditions and 36 rate constants. Our preliminary simulation results revealed no significant differences in the long-term dynamics between model outputs with or without pre-equilibration (data not shown). Thus, to reduce the running time, the model was not pre-equilibrated before running the simulations.

We measured an extensive set of experimental data for parameter estimation because many of the non-zero initial conditions and the rate constants for the Fas pathway are unknown. Moreover, the initial conditions and the rate constants for the aggregated molecules do not correspond to any values in real life. Cleavage kinetics of full-length caspase-8 with stimulation of 100 ng ml^{-1} FasL in control Jurkat cells are the same data that have been described previously (Hua *et al.* 2005). Caspase-3 cleavage kinetics in both control (transfected with empty vector) and Bcl-2 overexpressed Jurkat cells (generation of these cell lines has been described previously (Hua *et al.* 2005)) stimulated with 1, 10, 100 and 1000 ng ml^{-1} FasL were measured using intracellular staining followed by flow cytometry (figure 2). All the experiments were repeated at least three times and gave similar results. Results from the best experiment were chosen for parameter estimation.

In order to fit the estimated parameter values to experimental data, the unknown initial conditions were manually adjusted (initial concentrations taken from literature are indicated in electronic supplementary material) and all the rate constants were optimized using Jacobian with the WEIGHTED_LEAST-SQUARES option as the objective function. The sum of free cleaved caspase-3 and XIAP bound cleaved caspase-3 was used to fit the FACS data of caspase-3 activation since the intracellular staining technique does not differentiate the two forms of cleaved caspase-3.

Since, a previous study (Hua *et al.* 2005) has shown that a sixfold increase of Bcl-2 level has the same blocking effect as a 50 fold increase, the initial condition for Bcl-2 was increased by 10 fold during

model simulation to mimic the experimental Bcl-2 overexpressed cells. All of the biochemical reactions, reaction rate constants and non-zero initial conditions used in the model are listed in electronic supplementary material. These values are referred to as baseline values. Sensitivity analysis was done in Jacobian using a non-normalized sensitivity calculation.

4.4. Monte Carlo simulation

Monte Carlo simulation was conducted in MATLAB with 'ode15s' differential equation solver. Each changing initial condition has a concentration range from 10% (10^{-1}) of to 10 times the baseline value. The sampling interval within this range was set to be $10^{0.025}$ unless otherwise specified. With this sampling interval, each initial condition has 81 potential values, resulting in $9^{81}=2.0\times10^{77}$ possible combinations of different initial conditions. Each run of Monte Carlo simulation contains one million individual simulations, each of which simulates 6 h of reactions after adding FasL. For each simulation, the initial conditions for all the molecules and the time-course of the free cleaved caspase-3 generation were saved for the following analyses.

4.5. Clustering

One million simulations from a run of Monte Carlo simulation were clustered into three groups according to the kinetics of free cleaved caspase-3 production using *k*-means clustering algorithm (Duda *et al.* 2001). We used the squared Euclidean metric as the distance metric to group together free cleaved caspase-3 time-courses with similar dynamics as well as similar final concentrations at the end of the 6 h simulation period. The *k*-means clustering algorithm was run five times and each run consisted of 500 iterations. The number of simulations in each cluster was 360 182 (sensitive-response cluster), 247 195 (medium sensitive-response cluster) and 392 623 (insensitive-response cluster). The centroids for all three clusters are given in figure 3.

4.6. Construction of the decision tree model

Decision trees are a family of algorithms that aim to uncover the predictive structure of a classification or prediction problem while still maintaining good prediction accuracy. The decision tree algorithm used in our study is the classification and regression trees (CART) method (Breiman *et al.* 1984). Here, we used initial conditions for each simulation as the predictor dataset ($\boldsymbol{X}\in\mathbb{R}^{1\,000\,000\times9}$), and corresponding clusters (computed in §4.5) as the outcomes ($\boldsymbol{Y}\in\mathbb{R}^{1\,000\,000\times1}$).

The idea behind the CART algorithm is to progressively split the predictor dataset into smaller and smaller subsets. Ideally, each subset would correspond to only one outcome, i.e. each subset would be *pure*. Usually, however, there is a mixture of outcomes after a split, so for each split there is a need to decide whether to continue splitting or accept imperfect classification. If splitting is chosen, all variables in the predictor dataset are considered and the one minimizing impurity is used to split the data subset again into two subsets. For example, in figure 4, the first split is done based on the initial condition of XIAP so that the dataset is divided into two subsets: cases with XIAP initial value less than 3.1 go to left ($\boldsymbol{X}_{\mathrm{XIAP}<3.1}\in\mathbb{R}^{740\,495\times9}$), while the rest go to right $\left(\boldsymbol{X}_{\mathrm{XIAP}\geq3.1}\in\mathbb{R}^{259\,505\times9}\right)$.

The original dataset is split until a termination condition (e.g. validation error is minimized, purity after a split does not increase significantly, or split dataset would have too few cases) is met. In order to avoid overfitting, the final decision tree is achieved by applying a pruning algorithm that aims at to simultaneously maximize the prediction accuracy and minimize the tree complexity (i.e. size). Here, we used the cost-complexity pruning algorithm as described in (Breiman *et al.* 1984).

Leaf nodes in a decision tree represent outcomes (one of the apoptosis sensitivity clusters). A rule in the decision tree is the path from the root node to a leaf node. For example, the rule I_1 (figure 4) is a result from two splits; first XIAP$\geq$3.1 and then XIAP$\geq$5.2, giving the rule '`IF` XIAP$\geq$5.2 `THEN` the class is insensitive response cluster.' Accuracy of this rule is 78%.

The parameters used in the CART algorithm were as follows. The Gini index (Breiman *et al.* 1984) was chosen to be the purity function instead of twoing and deviance functions since it gave slightly better prediction accuracy than the other two functions. In order to avoid leaf nodes containing too few data points, we constrained the number of simulations in a node to be at least 15% of the number of simulations in the smallest cluster for the node to be split again. In our case, each node must contain at least 37 080 simulations in order to be considered for splitting. Prior probability for *i*th class was obtained by dividing the number of the cases of *i*th class by the total amount of observations. The costs of a misclassification and a correct classification were set to be one and zero, respectively.

4.7. Calculation of the COST values

In order to estimate the minimal number of perturbations between any two clusters, we define COST value to be the estimated number of molecules with different concentration ranges between two paths. These are the molecules need to be perturbed to switch a cell from one path to another path. The COST for switching from path X (e.g. I_2) to path Y (e.g. III_1) is calculated in the following way:

(i) `IF` one molecule belongs to both paths and the concentration range of this molecule for path X is a subset of the range for path Y, `THEN` COST is unchanged;
(ii) `IF` one molecule belongs to both paths and the concentration range of this molecule for path X has no intersection with the range for path Y, `THEN` COST is increased by one;
(iii) `IF` one molecule belongs to both paths and the concentration range of this molecule for path X partially overlaps with the range for path Y, `THEN` we estimate the likelihood that this molecule has different value between path X

and Y by calculating the percentage of simulations in path X with the concentration value of this molecule not in the range for path Y and COST is increased by this percentage;

(iv) IF one molecule belongs to path Y, but not to path X, THEN COST is increased by the percentage of simulations in path X with the concentration of this molecule not in the range for path Y and

(v) IF one molecule belongs to path X, but not to path Y, THEN COST is unchanged.

An example of a COST value calculation is given in the electronic supplementary material.

We would like to thank Dr Suzanne Gaudet, Dr Melissa Kemp and Lucia Wille for critically reading the manuscript and for helpful comments and insights. This work was supported by an NIGMS P50 grant to the MIT Cell Decision Processes Centre, by a gift from Entelos, by the Academy of Finland and Helsingin Sanomain 100-Vuotissäätiö.

REFERENCES

Adam, B. L. *et al.* 2002 Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.* **62**, 3609–3614.

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R. & Califano, A. 2005 Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* **37**, 382–390. (doi:10.1038/ng1532)

Bentele, M., Lavrik, I., Ulrich, M., Stosser, S., Heermann, D. W., Kalthoff, H., Krammer, P. H. & Eils, R. 2004 Mathematical modelling reveals threshold mechanism in CD95-induced apoptosis. *J. Cell. Biol.* **166**, 839–851. (doi:10.1083/jcb.200404158)

Bratton, S. B., Lewis, J., Butterworth, M., Duckett, C. S. & Cohen, G. M. 2002 XIAP inhibition of caspase-3 preserves its association with the Apaf-1 apoptosome and prevents CD95- and Bax-induced apoptosis. *Cell Death Differ.* **9**, 881–892. (doi:10.1038/sj.cdd.4401069)

Breiman, L. 2001*a* Random forests. *Mach. Learn.* **45**, 5–32. (doi:10.1023/A:1010933404324)

Breiman, L. 2001*b* Statistical modelling: the two cultures. *Stat. Sci.* **16**, 199–231. (doi:10.1214/ss/1009213726)

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. 1984 *Classification and regression trees*. London: Chapman & Hall.

de Pillis, L. G., Radunskaya, A. E. & Wiseman, C. L. 2005 A validated mathematical model of cell-mediated immune response to tumor growth. *Cancer Res.* **65**, 7950–7958.

Duda, R. O., Hart, P. E. & Stork, D. G. 2001 *Pattern recognition*. London: Wiley.

Frantz, S. 2005 Drug discovery: playing dirty. *Nature* **437**, 942–943. (doi:10.1038/437942a)

Fussenegger, M., Bailey, J. E. & Varner, J. 2000 A mathematical model of caspase function in apoptosis. *Nat. Biotechnol.* **18**, 768–774. (doi:10.1038/81208)

Gunther, E. C., Stone, D. J., Gerwien, R. W., Bento, P. & Heyes, M. P. 2003 Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. *Proc. Natl Acad. Sci. USA* **100**, 9608–9613. (doi:10.1073/pnas.1632587100)

Hautaniemi, S., Kharait, S., Iwabu, A., Wells, A. & Lauffenburger, D. A. 2005 Modelling of signal-response cascades using decision tree analysis. *Bioinformatics* **21**, 2027–2035. (doi:10.1093/bioinformatics/bti278)

Hoffmann, A., Levchenko, A., Scott, M. L. & Baltimore, D. 2002 The IkappaB–NF-kappaB signalling module: temporal control and selective gene activation. *Science* **298**, 1241–1245. (doi:10.1126/science.1071914)

Hua, F., Cornejo, M. G., Cardone, M. H., Stokes, C. L. & Lauffenburger, D. A. 2005 Effects of Bcl-2 levels on Fas signalling-induced caspase-3 activation: molecular genetic tests of computational model predictions. *J. Immunol.* **175**, 985–995.

Igney, F. H. & Krammer, P. H. 2002 Death and anti-death: tumour resistance to apoptosis. *Nat. Rev. Cancer* **2**, 277–288. (doi:10.1038/nrc776)

Irmler, M. *et al.* 1997 Inhibition of death receptor signals by cellular FLIP. *Nature* **388**, 190–195. (doi:10.1038/40657)

Kholodenko, B. N., Demin, O. V., Moehren, G. & Hoek, J. B. 1999 Quantification of short term signalling by the epidermal growth factor receptor. *J. Biol. Chem.* **274**, 30 169–30 181. (doi:10.1074/jbc.274.42.30169)

Kislinger, T., Gramolini, A. O., Pan, Y., Rahman, K., MacLennan, D. H. & Emili, A. 2005 Proteome dynamics during C2C12 myoblast differentiation. *Mol. Cell Proteomics* **4**, 887–901. (doi:10.1074/mcp.M400182-MCP200)

Krammer, P. H. 2000 CD95's deadly mission in the immune system. *Nature* **407**, 789–795. (doi:10.1038/35037728)

Landowski, T. H., Qu, N., Buyuksal, I., Painter, J. S. & Dalton, W. S. 1997 Mutations in the Fas antigen in patients with multiple myeloma. *Blood* **90**, 4266–4270.

Lavrik, I., Golks, A. & Krammer, P. H. 2005 Death receptor signalling. *J. Cell Sci.* **2**, 265–267. (doi:10.1242/jcs.01610)

Ma'ayan, A., Blitzer, R. D. & Iyengar, R. 2005 Toward predictive models of mammalian cells. *Annu. Rev. Biophys. Biomol. Struct.* **34**, 319–349. (doi:10.1146/annurev.biophys.34.040204.144415)

Mencher, S. K. & Wang, L. G. 2005 Promiscuous drugs compared to selective drugs (promiscuity can be a virtue). *BMC Clin. Pharmacol.* **5**, 3. (doi:10.1186/1472-6904-5-3)

Nagata, S. 1999 Fas ligand-induced apoptosis. *Annu. Rev. Genet.* **33**, 29–55. (doi:10.1146/annurev.genet.33.1.29)

Podgorelec, V., Kokol, P., Stiglic, B. & Rozman, I. 2002 Decision trees: an overview and their use in medicine. *J. Med. Syst.* **26**, 445–463. (doi:10.1023/A:1016409317640)

Rosenwald, A. *et al.* 1947 The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.* **346**, 1937–1937. (doi:10.1056/NEJMoa012914)

Roth, B. L., Sheffler, D. J. & Kroeze, W. K. 2004 Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.* **3**, 353–359. (doi:10.1038/nrd1346)

Scaffidi, C., Fulda, S., Srinivasan, A., Friesen, C., Li, F., Tomaselli, K. J., Debatin, K. M., Krammer, P. H. & Peter, M. E. 1998 Two CD95 (APO-1/Fas) signalling pathways. *EMBO J.* **17**, 1675–1687. (doi:10.1093/emboj/17.6.1675)

Schoeberl, B., Eichler-Jonsson, C., Gilles, E. D. & Muller, G. 2002 Computational modelling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat. Biotechnol.* **20**, 370–375. (doi:10.1038/nbt0402-370)

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. & Friedman, N. 2003 Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176.

Wiley, H. S., Shvartsman, S. Y. & Lauffenburger, D. A. 2003 Computational modelling of the EGF-receptor system: a paradigm for systems biology. *Trends Cell Biol.* **13**, 43–50. (doi:10.1016/S0962-8924(02)00009-0)