Elucidating the effects of SNPs in genotype-phenotype mappings of carcinogenesis

Jan Erik Jimmie Weiss

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

The University of Sheffield
Faculty of Medicine, Dentistry & Health
Department of Oncology and Metabolism

July 2019

Acknowledgements

Let me start by thanking the University of Sheffield and the Insigneo Institute for funding this journey of mine and my two supervisors, Prof. Angela Cox and Prof. Nick Monk, for supporting me all the way through. It has been a pleasure to run between the School of Mathematics and Statistics and the Medical School with my whimsical ideas, always feeling encouraged to pursue them. It is said that you learn from your mistakes and I have truly learned a lot over these years. I have enjoyed every single one of our meetings and all the diversions into philosophy and everything off topic they have taken. I cannot recall coming out of a single meeting, feeling like I knew what I was doing, but I was always confident that is the way it is supposed to be when you are doing research, and that we were on the right track.

My thanks also go to all the people I have had the honour of sharing office with over the years. I have had a great time both in and outside of the office and you have made this time so much more enjoyable.

Naturally I also want to thank my parents, without whom I would not be where I am today. You have always encouraged me to explore and try new things and supported me in my endeavours.

Last, but definitely not least I want to thank my wonderful wife Eva Weiss and my two lovely children, Thor and Loki. You have always been by my side, reminding me of what is important when my priorities strayed, and supporting me in the darkest hours, when it looked as if all the hard work had been for naught. This has been your journey as much as it has been mine. I am truly grateful for the smiles, hugs and kisses to pull me up when things were tough and the screams and cries to ground me when I was lost in thoughts, or just selfish enough to want a few consecutive hours of sleep.

Abstract

Germline genetic variations have been shown to affect the overall risk of developing cancer. In this thesis I combine mathematical modelling of gene regulatory cell and signalling pathways with genetic and molecular data with the aim to gain understanding of the mechanistics behind this association.

I started by evaluating the suitability of standard sensitivity analysis tools to study the link between risk associated genotypes and model dynamics corresponding to phenotype changes linked to carcinogenesis. From the sensitivity analysis it became clear that, although the development from normal tissue to cancer is gradual, on a dynamics level, the parameter space of the model could be divided into a more or less binary state space representing healthy and diseased phenotype.

Using this insight, a novel method was developed for studying how dynamical changes caused by a genotype effect its link to the risk of developing cancer. This method was built on the hypothesis that the distance between the initial location in parameter space and the border between the two phenotypes could be used as a proxy for the risk of developing cancer. The method was evaluated using theoretical data and it was shown that both the dynamics of the model and the results from the new framework correlated strongly with the relative risk attributed to the genotypes, even when noise was introduced into the underlying data.

The developed method was applied to two cancer types and three different cell samples. The results from this analysis were inconclusive, which, by looking back at the theoretical analysis, could partially be explained by the small sample sizes. Nevertheless, the theoretical, in combination with the experimental results, indicate that the framework proposed in this thesis could be used to bridge the gap between the molecular dynamics and the genetics of carcinogenesis.

Contents

List of Figures

List of Tables

Abbreviations

ABM Agent-Based Model.

eQTL Expression Quantitative Trait Locus.

FDR False Discovery Rate.

GO Gene Ontology.

GWAS Genome Wide Association Study.

IC initial condition.

KEGG Kyoto Encyclopedia of Genes and Genomes.

LD linkage disequilibrium.

ODE Ordinary Differential Equation.

PC Principal Component.

PCA Principal Component Analysis.

PDE Partial Differential Equation.

PSS Parameter Sensitivity Spectrum.

RPKM Reads Per Kilobase (of transcript) per Million (of reads).

SHM Sensitivity Heat Map.

SNP Single Nucleotide Polymorphism.

SS steady state.

STRING Search Tool for the Retrieval of Interacting Genes/Proteins.

SVD singular value decomposition.

TCGA The Cancer Genome Atlas.

Chapter 1

Biology of Cancer and the Use of Computational Modelling to Gain Deeper Understanding

As knowledge about the cellular system has expanded it has become increasingly clear that cellular signals travel in distinct pathways interacting with each other to create a regulatory network. Over the years, the map of many parts of the network has become so detailed and complex that it is impossible to comprehend how a change in the network will propagate and affect other parts of the system. However, it is possible to study the dynamics of constrained pathways using computational modelling tools. These models have been able to give new insight into the complex system and guide researchers into choosing experiments with high likelihood of giving new biological knowledge. By introducing perturbations into these models they have also been used extensively in the pursuit of extended knowledge about cancer development and ways of treating cancer.

At the same time, the means of acquiring molecular data on cellular signalling systems are becoming increasingly high-throughput and in order to be able to analyse and interpret the vast amount of data computational methods have become ever more sophisticated. Through this development of data acquisition and analysis tools it has been possible to identify hereditary genetic variations with very small, but statistically significant associations with cancer susceptibility.

Cancer is not only already one of the leading causes of death according to the World Health Organization but the number of new cases and the amount of people in need of chemotherapy are also expected to rise by around 50% by 2040 [1, 2]. Better diagnostics and treatments are still needed and will be even more so in the future. Applying the new knowledge derived from large molecular datasets within computational modelling opens up the possibility to study the molecular dynamics leading to cancer susceptibility

and gain knowledge hidden behind complexities that experimental methods alone cannot breach.

1.1 Cancer signalling pathways

Cancer as a disease has been studied for many decades. Even though many details of carcinogenesis are still to be unravelled, the broader picture was very well illustrated in the 6 hallmarks of cancer by Hanahan and Weinberg (2000) [3] (Figure 1.1). These were later complemented with two further hallmarks and two enabling characteristics in 2011 [4]. Not surprisingly, many of these hallmarks work in similar or complementing pathways in the cell. Furthermore, these pathways are often enriched in tumour suppressor genes or oncogenes such as TP53, RB, PIK3CA and FGFR-1 [5]. Some examples of important cancer pathways are summarized below.

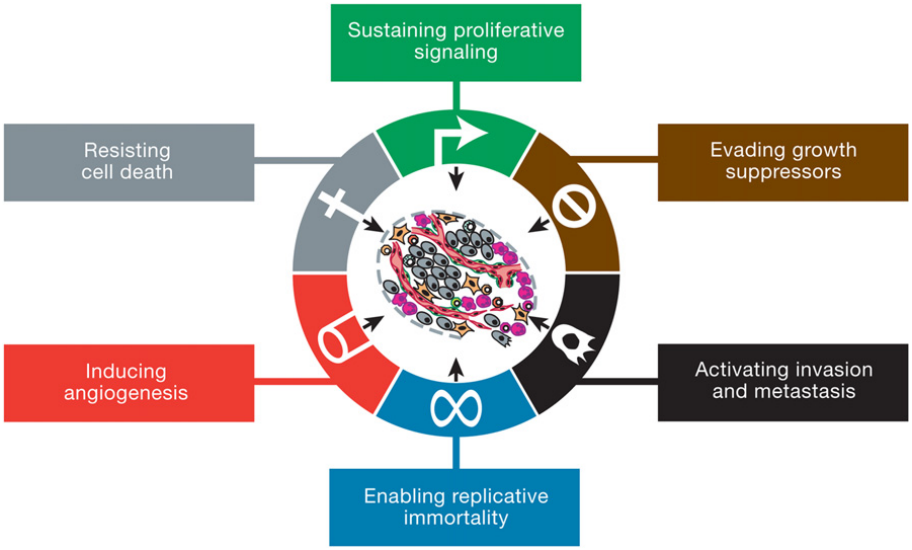


Figure 1.1: The hallmarks of cancer as first suggested by Hanahan et al. in 2000. In addition to these six hallmarks another two hallmarks were later suggested, these being: deregulating cellular energetics; and avoiding immune destruction. Furthermore, two enabling characteristics were suggested: genome instability and mutation; and tumour-promoting inflammation. Reprinted from Cell, Vol 144, Douglas Hanahan and Robert A. Weinberg, Hallmarks of Cancer: The Next Generation, 646-674, 2011, with permission from Elsevier.

1.1.1 Growth and growth suppressor signalling

For a cell to grow and proliferate, it is often dependent on growth signals from the surroundings. These interact with growth signal receptors on the cell surface, which transfer

the signal into the cell and propagate it in designated pathways. A cell achieves sustained proliferation signalling, one of the hallmarks of cancer, by producing its own growth factors/hormones, or by regulating the intracellular signalling, commonly through mutations in key genes [4]. To exemplify, a single point mutation in BRAF2 of the Ras-Raf-MEK-ERK pathway could render the protein constitutively active, thus decreasing the dependence on upstream activation by extracellular growth factors [6]. Similarly a mutation in PIK3CA, as well as mutations in up or down stream components, could increase the activation of the cell proliferation PI3K pathway [7, 8].

Another example is found within the p53 pathway. TP53 is one of the most important tumour suppressor genes. Through p21 and together with Rb it controls the cell cycle by regulating Cyclin D and E2F activity [9]. Down-regulation of Rb is a common path to gaining resistance to growth suppressing signalling [10].

1.1.2 Death signalling and immortalization

As a means of protecting itself from over-expression of growth signalling and malfunctioning of the cell machinery, the cell can induce apoptosis, i.e. programmed cell death. This can happen through extracellular signalling via death signal receptors or intracellular signalling via stress response agents [11]. Both of these pathways activate a cascade of Caspases that in the end lead to death of the cell. The p53 response to DNA damage is a good example of such signalling and many times de-regulation of p53 also leads to the cell resisting apoptosis signals. However, this is also accomplished by up-regulation of anti-apoptotic factors such as Bcl-2 [11] or down-regulation of apoptotic factors such as FasL, BAK or BAX [12, 13, 14].

1.1.3 Angiogenesis

A limiting factor for further growth is the cell's ability to recruit blood vessels (angiogenesis) to ensure sufficient supply of oxygen and nutrients. Many cancers have the ability to induce angiogenesis through expression of vascular epidermal growth factors (VEGFs) [15, 16] or over-expression of cyclo-oxygenases (COXs), key regulators of VEGF expression [17, 18]. Unsurprisingly the capability to promote lymph angiogenesis is linked to many types of metastasising cancers [19, 18].

1.1.4 Motility and metastasis

A very severe trait of late stage cancers is their ability to escape their micro-environment and spread to other parts of the body, a process known as metastasis. This is often

accomplished by regulating the expression of cell adhesion molecules [4]. E-Cadherin is the main cell-cell adhesion molecule in epithelial cells and its regulation is linked to changes in motility and invasiveness of cancer cells [20, 21]. Additionally, up-regulation of N-cadherin has been shown to promote motility and invasiveness in various cancers [22, 23]. An example of the interconnectedness of the pathways is that N-cadherin, together with E-cadherin, also protects against apoptosis through activation of PI3K and Akt pathways [24, 25].

1.2 Genetics of cancer susceptibility

As our understanding of the mechanisms behind carcinogenesis has evolved questions have been raised about not only how cancer is developing but also what factors are driving this development. It has long been known that members of families with a history of cancer often run a greater risk to develop cancer themselves. The primary focus of this thesis will be breast cancer development and early work within this field did indeed identify the two genes BRCA1 and BRCA2 where individual genetic variations played a major role in the variation of breast cancer susceptibility [26, 27]. Both BRCA1 and BRCA2 are involved in DNA repair and mutations within these genes are linked to a 10- to 20-fold increase in the risk of developing breast cancer. Although no other genes with such an impact have been identified, there are some genes with rare to moderate prevalence of mutations in populations predisposed to developing certain types of cancer. Many of the genes involved in breast cancer susceptibility are interacting with BRCA1 or BRCA2, such as BRIP1, PALB2, ATM and CHEK2. The latter two are also regulating p53 response to DNA damage [28].

Recently, the emergence of newer technology has opened up the possibility to screen tens of thousands of patients for Single Nucleotide Polymorphisms (SNPs) on a genome wide scale. SNPs are common germ line single nucleotide variants with a prevalence of at least 1% within the population [29]. Using a methodology called Genome Wide Association Study (GWAS) which compares the genotypes of groups of individuals with and without a given phenotype, or disease, makes is possible to find correlations between the SNPs and these phenotypes or susceptibility to these diseases. Even though these alleles are linked to a very limited fold-increase in cancer susceptibility, worldwide collaborations, providing large sample sizes, have made it possible to establish an ever-growing list of susceptibility loci. As a result SNPs can now explain 14% of breast cancer susceptibility [30]. While many of these alleles are located in regions with no known function and several Mb from the closest known gene [31, 30], many also are located close to, but not

within known protein coding regions with links to pathways involved in carcinogenesis. In fact, among 41 breast cancer associated SNPs characterised in 2013, 19 were intergenic and 20 were intronic, whereas 2 fell within protein coding regions [30]. This pattern can be seen in other cancer associated SNPs as well. Among 63 prostate cancer associated SNPs 32 were located within introns and 2 in the 3'-untranslated regions [32]. Only four of the SNPs were missense-variants. Some of the genes with SNPs linked to breast cancer susceptibility include the apoptotic inducer CASP8 [33], the growth factor receptor and growth factor binding protein FGFR2 [34] and IGFBP1 [35] and the angiogenic factor VEGF. Other examples are genes in similar pathways such as MAP3K1 [36] and ESR1 [37], and the telomerase prolonging TERT [38].

However, since individual susceptibility loci are relatively common in the general population and have very little effect on overall cancer susceptibility, they are very difficult to identify. Even with very large sample numbers results can be difficult to replicate. This is illustrated by the case of a CASP8 SNPs, which showed a much weaker link to cancer susceptibility in a later study than in the initial study [39, 33].

Overall the human genome has been estimated to contain around 7 million common SNPs (>5% prevalence) [40], too many to be genotyped in large scale studies. However, many of these SNPs are located in close proximity to each other, in which case their genotypes tend to be highly correlated with one another, and they are said to be in linkage disequilibrium (LD). It is therefore common to only select one or a few SNPs from each cluster and treat them as a proxy for the neighbouring SNPs in that region. By doing that, association studies of the entire genome can be performed using only a few hundred thousand SNPs [41]. This comes with the drawback that one cannot be certain whether the observed association to a phenotype is linked directly to the genotyped allele or to a nearby SNP that is highly correlated to the one included in the study.

In order to be able to identify the actual causal polymorphisms a fine-scale mapping of the region around an interesting disease-associated SNP is often necessary. Indeed, a fine-scale mapping of the region around CASP8 found four significant SNPs located in or around CASP8, out of which one could be replicated in a larger follow up study [33]. Similarly, by investigating approximately 480 SNPs around the TERT gene and including data of telomere length, two distinct SNPs were linked to increased risk of various cancers due to elongated telomeres, and two additional SNPs were linked to cancer due to generation of a truncated dysfunctional TERT [38]. A third example involved the fine-scale mapping of the region around FGFR2, which identified two SNPs in transcription factor FOXA1- and E2F-binding regions of FGFR2 [42]. Data such as this and the fact that cancer-associated SNPs are often found in gene regulatory regions indicates that SNPs

could regulate gene expression. Indeed, many SNPs have been shown to act as Expression Quantitative Trait Loci (eQTLs), loci associated with a change in expression of a gene [43, 44]. These eQTLs regulate gene expression, not only of nearby genes, but sometimes genes on other chromosomes as well.

These three cases highlight the possibility to pinpoint specific disease susceptibility alleles and link them to a biological function. As of 2017, over 170 loci had been reported to be associated with breast cancer susceptibility and over one thousand additional polymorphisms have been estimated to affect breast cancer susceptibility [30, 45]. By searching for pathways enriched in SNP targets [46], or identifying transcriptome level changes correlated to polymorphisms [47, 48], the biological function of SNPs linked to various types of cancer have started to be unravelled. These efforts, in combination with more fine-scale mapping, will increase the number of alleles with known biological consequences. This raises the possibility of incorporating the functional effects of SNPs into computer models of cellular signalling pathways, to study their role in cancer development.

1.3 Mathematical Modelling

Over time a wide range of modelling techniques have been developed, each with their own possibilities and drawbacks. When trying to use mathematical modelling as a tool to answer a specific question it is important to be aware of these strengths and weaknesses and to choose a modelling technique suitable for the question in mind. One way of classifying models would be by how they treat interactions within the network. Roughly, most models can be categorised as either logical or continuous models, agent-based models or hybrids thereof. All of these classes of modelling have their advantages and disadvantages, as will be seen in the following sections.

Even if it is not essential to know the precise function of a SNP in order to model its effect, the more accurate the knowledge is, the more accurately it can be represented in the model. As will be seen later, it also limits the extent to which different modelling methods can be used.

1.3.1 Logical Models

The simplest form of a logical network, and one of the first models of regulatory networks that were proposed [49, 50], is a so called Boolean network. It consists of a directed graph, where each compound or observation is represented by a vertex (or a node) and the interactions between compounds are edges connecting the vertices. The model has

Boolean updating rules, where each node can be either ON or OFF and during updating, the rule takes into account the state of each of the incoming edges (Figure 1.2).

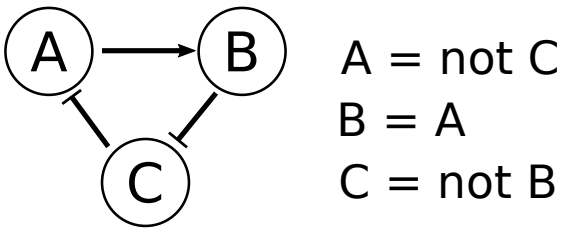


Figure 1.2: Schematic of a simple boolean model. A is inactivated by C, B is activated by A and C in inactivated by B.

Even though binary states and the discrete time scale that results from the updating scheme pose potential problems for a Boolean model, they also act as the main advantages. A Boolean model does not require any parameters for all the interactions, but instead focuses on the core regulatory logic. At the same time the state space (possible combinations of states) is finite, thus making it possible to exhaustively investigate the possibilities within the system. A Boolean model is therefore a good first start when not much is known about the details of the interactions in the network. Furthermore, when the Gene Regulatory Network (GRN) studied is highly dependent on the topology of the network, it is possible to achieve the same results using a Boolean approach as a continuous modelling mentioned later [51, 52].

The simplicity of the modelling approach also allows for very large networks to be modelled. As an example of this, a simple synchronous model implementing the core parts of carcinogenesis was constructed by Fumiã and Martins in 2013 [53]. By including 97 primary genes and external stimuli for cell development and survival the model was able to predict the effects of deletions and constitutive over-expressions of genes depending on environmental circumstances. By sequentially altering the expression of 7 genes they could see how the cell lost sensitivity to death stimuli and showed increased signs of cell immortality. Moreover, they were able to estimate the severity of the progression in various environments, such as normoxia vs. hypoxia or sufficient vs. insufficient nutrient supply. Finally they were able to assess the efficiency of mono vs. pluri-therapies in suppressing the carcinogenic traits of the cell [53].

1.3.2 Continuous models

Both the strength and the weakness of the logical model lie in its simplicity. By considering an edge as either active or not, the model fails to take into account the temporal dynamics

of interaction intensities. Continuous models provide a better alternative to model the temporal dynamics of a system, and one of the most common types of continuous models consist of a set of Ordinary Differential Equations (ODEs) [54, 55, 56, 57].

These models are generally written in the form:

$$\frac{dy}{dt} = f(y, k), \qquad y(0) = y^0 \tag{1.1}$$

where y is a vector of variables, k is a vector of parameters linked to these variables, and y^0 are the initial conditions (at time=0).

The complexity of GRNs and the need for experimental data to reliably determine the parameters of the model greatly limits the extent to which one can usefully model larger systems using differential equations (DEs). DE-modelling is, nevertheless, a very powerful tool to gain deeper insight into experimentally well-studied systems such as the cell cycle. Logical models are limited to studying inactivating or constitutive- activation-rendering mutations. In contrast, models such as the ones by Novák and Tyson (2004) [58] or Csikász-Nagyet al. (2006) [55] are able to investigate the effect of smaller, gradual perturbations in the system on the outcome of the cycle. As with any computational work, there is a risk of over-interpreting the data and as Weis et al. (2014) [57] showed, when comparing the previously mentioned models, a model is only a rough approximation of the reality and each model has its limits.

Even though the models are just approximations of the real system, using ODEs it is possible to model intra- and intercellular reactions at the same time as was shown by Jain et al. (2008) [59], where the VEGF-Bcl-2-CXCL8 pathway involved in angiogenesis was modelled. In this work Bcl-2, which is also involved in apoptotic signalling, was recognized as a potential drug target to inhibit cancer progression and spread. This was later tested in another model focusing on the Bcl-2 response to an inhibiting drug [60]. Even though the latter model was very simplified, it reproduced the results from *in vitro* studies sufficiently and predicted a relatively poor performance of the drug in question *in vivo*. Another example of an application of these methods is the FIH-PHD-HIF pathway model by Nguyen et al. (2013) [61], which not only accurately represented the HIF-1a regulation of HRE genes, but also gave insights and likely explanations to seemingly counter-intuitive experimental results.

1.3.3 Agent Based Models

In addition to the logical and the continues models, another large group of models which are used extensively in studying cancer are Agent-Based Models (ABMs). These models

can simulate interactions and fates of agents at various scales by letting the modelled features be governed by a set of rules (much like the logical models, but the rules are not limited to logical rules). Using this framework entire tissues can be modelled by treating each cell as an agent. The actions of each cell can be simulated based on which pathways are activated, how it interacts with surrounding cells or how its micro-environment looks, in terms of nutrients, oxygen or other important factors. By conceptualising factors in the model, each factor can be modelled on an appropriate level, resulting in multi-scale models. For example, individual pathways can be modelled using ODEs, and the activation of these pathways can result in different cell behaviour depending on a set a rules. These rules can also depend on larger scale factors, such as nutrient diffusion, often modelled using Partial Differential Equations (PDEs). The spacial aspect of these models render them particularly suitable to study phenomena such as tumour growth [62] and cell migration [63]. They have also been used to study angiogenesis, vascularisation and invasiveness [64].

To some extent ABMs have also been used to study the role of the genome in cancer development. Gerlee and Anderson used a hybrid cellular automata model (an approach related to ABM) to study clonal evolution in cancer and under what circumstances the glycolytic phenotype associated with tumours is most likely to arise [65]. Araujo et al. [66] used an ABM to study the effect of chromosome missegregation (when chromosomes get unequally split between daughter cells during cell division) on cancer development. Anderson et al. also studied the evolution of cancer phenotypes through mutations and how they were selected for under pressure from the micro-environment [67].

Although these models are very powerful, they are also very computationally intense and often achieve their large scale and high through-put capabilities by abstracting away much of the details in the underlying pathways. This trade-off may be acceptable if the centre of attention is a phenomenon occurring at a larger scale. However, it makes them less suitable to the study the details of the link between the genotype and the phenotype.

1.3.4 Mathematical models of cancer development and fitness landscapes

Another approach to model cancer development is based on the concept of carcinogenesis as a process of evolution. The cancer cells are viewed as occupying a fitness landscape where different genotypes are associated with different levels of fitness. As the cell accumulates mutations it moves in genotype space and acquires different levels of fitness. If the new mutation results in the cell occupying a space of higher levels of fitness it has

acquired an evolutionary advantage over surrounding cells which will cause it to expand within the tumour relative to the other cells.

The path a cell takes in this fitness landscape as it evolves into a more malignant cancer can be seen as an evolutionary trajectory through the fitness landscape. A big question which has been debated a lot is to what extent these evolutionary trajectories are predictable. To date there are studies which point both towards examples where this, is the case as well as cases where it is not. For example, G. Caravagna *et al.* applied a machine learning method based on transfer learning to a large dataset of multi-region sequencing of tumours to infer evolutionary patterns across patients. Using this method they could identify several recurrent evolutionary trajectories in lung, breast and renal cancer[68].

A problem with the approach of modelling the fitness landscape is that it relies on genotype-phenotype mappings which then in turn map onto the fitness landscape. In practice it is very difficult to measure fitness, especially in *in vivo* systems. Consequently, a lot of work has been based on simulated data.

R. Diaz-Uriarte used statistical cancer progression models to represent feasible trajectories in the fitness landscape due to evolutionary constraints as directional acyclic graphs (DAGs)[69]. He then simulated the evolution on various fitness landscapes using a continuous-time, logistic-like model where cells stochastically divide and acquire mutations which move them in the fitness landscape, making them more or less likely to survive long enough to divide again. When applying the method to three different cancer data sets he could show that the level of predictability degraded with the presence of reciprocal sign epistasis (when two or more genotypes have a negative effect if observed on their own, but a positive effect when observed together); a phenomenon believed to be important and common in cancer development. Widely different fitness landscapes could give rise to genotype frequencies empirically observed in the cancer types, which then manifested in widely different DAGs with different constraints.

S-R Hosseini *et al.* on the other hand used a similar method based on conjunctive Bayesian networks (CBNs) to circumvent the need to measure fitness changes caused by mutations[70]. Using this method they were able to investigate the predictability of cancer evolution from mutational data and could show that under the assumption of strong selection and weak mutation rate, there was a strong correlation between the predictability of the CBN model and that of classical fitness landscape models. When applying the method to 15 different cancer types and considering a small number of frequent driver genes, they could show that many cancers had a high level of predictability with few evolutionary trajectories being likely.

One obstacle towards evolutionary predictability is the large amount of intratumoural heterogeneity seen in many cancers and many studies suggest that the predictability degrades quickly beyond a few strong driver mutations. For example M. Williams *et al.* used a mathematical model of the accumulation of mutations in a tumour to estimate the relative presence of mutations between cells under different conditions[71]. By relating the results from these models to sequencing data of tumour samples they could show that the heterogeneity seen in many cancers of different types was due to neutral evolution, meaning that most of the mutations causing clonal selection occurred before the onset of cancer growth.

Likewise, Sottoriva *et al.* suggested and validated a 'Big Bang' model of colorectal cancer growth in which the driver mutations occur early in the development and the tumour thereafter grows, mainly as a single clonal expansion with very little selective pressure[72]. This results in a tumour with a high level of intratumoural heterogeneity where most of the common alterations occur early after the expansion.

1.4 How does the genetics connect to the mathematical model

When creating a mathematical dynamical model for a pathway there are three things which connect the model to the biological pathway. The first is the topology of the network. In a biological network certain proteins interact with each other, and for the model to be able to represent the network it needs to represent these interactions in its own wiring. The two other important aspects of the network are the expression levels of the proteins and the interaction strengths and reaction speeds. Depending on the model, the levels of the proteins can be modelled either as initial conditions of the variable values or as a combination of the initial conditions and parameters representing production and degradation of the proteins. All of the interaction strengths and reaction speeds are set by parameter values of the equations of the involved proteins.

If all of the above characteristics of the pathway are included in the model, then any genetic variation altering the function or expression of a protein can be modelled by altering the initial conditions and/or the parameters representing these altered aspects of the pathway. In the case of a Boolean model, any mutations rendering a protein inactive or constitutively active can be modelled by altering the rules of the model to represent constant inactivation or activation. To some extent under- and over-expressions can be modelled in a similar way. A good example of such an approach being successful is the modelling of carcinogenesis by Fumiã and Martins [53] previously mentioned in section

1.3.1. In a continuous model these genetic alterations can be modelled with a much higher resolution by treating the mutations as perturbations of the initial conditions or parameters in the model.

What all of these methods have in common is that they consider what perturbations of the model will break it, i.e. what perturbation will cause its output to no longer represent the typical behaviour of the system. When considering the risk of developing a disease it is not the effects the genetic variations have on the output of the system that is important, but how that affects the risk of the system to break later on. The risk can therefore be seen as the potential of a quantitative change in the output to result in a qualitative change, given additional perturbations. It connects the sensitivity of the system (to what extend a perturbation changes the quantity of the output) to the robustness of the system (how easily a perturbation changes the quality of the output) and under what circumstances the former affects the latter.

To better understand the way risk is thought of in this work, one can consider two persons driving to work as an analogy of the biological system being modelled. Both persons have individual circumstances affecting their ability to drive and both cars are in good standard, but not identical to each other. For example, one person might be a morning person, whereas the other person is not, resulting in one of them needing a longer time to react than the other (given that they get the same amount of sleep). These conditions can be seen as the genotypes affecting a biological pathway and the initial conditions of a model. Under normal circumstance both drivers will every now and then encounter dangerous situation in traffic. However, since they are rested and focused and the cars are working properly, an accident will generally be avoided. This is the way the biological system works under normal circumstances. If the brakes stop functioning both will end up in an accident, regardless of other circumstances. This could represent major somatic mutations in the pathway. If a crucial part of the system breaks the pathway will not be able to perform its function. However, if the brakes only deteriorate a little bit the stopping distance will get longer, but if the driver is alert, he will still have enough time to stop before an accident occurs. This can represent somatic mutations with smaller effects which only alter the parameters or initial conditions slightly and consequently only affect the output of the model quantitatively, but not qualitatively. However if one driver happens to not be a morning person, he will be more tired than the other person and need a longer reaction time, resulting in him not being able to avoid the accident. In this case, circumstances which did not affect the outcome under normal circumstances (i.e. slight deterioration in the brakes), all of a sudden became crucial to determine the outcome. In the same way, a genetic variation may not affect the ability of the pathway to perform its

function in the cell under normal circumstance. However, upon accumulation of certain mutations the effect of the genetic variations become the difference between a normally behaving pathway and an abnormally behaving pathway.

The aim of this thesis is to develop a methodology to investigate and quantify changes in the risk of developing cancer in terms of the effect perturbations in a dynamical systems model have on the robustness of the output.

1.5 Thesis plan

Over the years certain pathways have crystallized as key components in carcinogenesis. These pathways are by no means complete and much is yet to be discovered, both when it comes to the components and their interactions, but also the way the different pathways are interacting and regulating each other. With the emergence of high throughput genotyping technologies it has also become possible to investigate the role of common genome variants and their effect on disease susceptibility. With the increasing amount of fine-scale mapping being performed, more and more SNPs are being associated with predicted biological functions.

With this in mind the aim of this thesis is to begin to answer the question:

What is the mechanism behind the effect carrying a particular SNP has on the risk of developing a particular disease phenotype?

The work will be based on the assumption that the effect an inherited genetic variation and the effect a somatic mutation has on the dynamics of the network can be modelled through perturbations of initial conditions and/or parameters in a dynamical systems model. Various computational models have already demonstrated their capability of predicting the outcome of several types of mutations on different levels of a cellular system, but very few models have incorporated SNP data. In fact, the vast majority of SNP-GRNs have been constructed using inference methods and are more concerned with the topology of the networks than studying the biological dynamics.

Since the SNPs change the risk of acquiring a new phenotype without severely changing the dynamics of the pathway, two key aspects of the work will be to: successfully link the effect of a SNP on dynamics of the biological pathway to the effect of a perturbation on the output of the model and; develop a methodology to relate the quantitative change a perturbation has on model dynamics to the risk of changing the output qualitatively.

By reasoning around their function, the biologically functional SNPs could be classified into regulatory changes of the kinds that logical or continuous models have been able to handle.

The challenges towards answering the main question of this thesis therefore will be to:

1. find a way to link SNPs to genes involved in the process of interest.
2. choose an appropriate mathematical model of that process.
3. decide for which aspects of that model the particular SNP corresponds to and how the SNP affects that aspect.
4. decide what features of the model will be used as a representation of the phenotype of interest.
5. decide on a method to assess the sensitivity of the chosen feature to changes in aspects of the model corresponding to the SNP.
6. use the outcome of the sensitivity to derive understanding regarding the mechanics of how the SNP is affecting the risk of the system to develop the disease phenotype.

These challenges will be addressed according to the following plan:

- In chapter two the SNPs associated with breast cancer and their connections to proteins which can be studied in known ODE models will be explored. In the end a set of ODE models will be chosen which will be the focus of the rest of the thesis.
- In chapter three, three conventional methods will be used to explore the sensitivity of the chosen models with regards to perturbations which could be the results of cancer associated SNPs. First, a standard parameter and variable scan will be used to map the sensitivities of the models from the initial position in state space. Then the sensitivity will be explored both at initial conditions representing normal cellular conditions and at various positions in variable/parameter-space which will represent the gradual development of a cell from normal to a cancer cell. This will be done using the sensitivity analysis tools SASSy and SloppyCell.
- In chapter four the lessons learned from the previous chapter will be used as a starting point to explore a novel way of measuring phenotype sensitivity. The concept of a phenotype separatrix will be introduced and the link to the biology on the one hand and the mathematics on the other will be discussed. A method for using this phenotype separatrix will be presented and applied to the two models explored in the previous chapter. The results from these models will be explored. Finally this method will be used on simulated data sets of cancer associated SNPs and the possibility to link the phenotype separatrix to the risk score of cancer associated SNPs will be explored.

- In the fifth and last experimental chapter the dynamics of one of the two models will be linked to experimental data from two tissue types, breast and prostate. The results from these simulations will be linked directly to the risk score ratio of relevant cancer associated SNPs. The risk score ratios will also be linked to the results from the phenotype separatrix analysis results obtained in chapter four, thereby correlating the risk of an individual to acquire cancer with the risk of a dynamical model to change output from normal to abnormal behaviour.

Chapter 2

Choosing the Models

2.1 Introduction

The previous chapter explained how the development of cancer works over a large range of pathways, whose control mechanisms gradually deteriorate, making the cell behave more and more like a cancer cell. We also saw how somatic mutations, acting cumulatively with inherited risk Single Nucleotide Polymorphisms (SNPs), contribute to the risk of associated pathways to break down rendering the individual more or less susceptible to cancer. Finally, in section 1.5, the aim of this thesis was declared to try to answer the question:

What is the mechanism behind the effect carrying a particular SNP has on the risk of developing a particular disease phenotype?

In the pursuit of answering this question a number of challenges were outlined. In this chapter the first two of these challenges will be addressed. First links between SNPs and genes in pathways of interest will be established. Then these links will be used as a basis for choosing a pathway and associated models to study in the next chapters.

2.1.1 Linking SNPs to genes and pathways

As of December 2018 there are 5,225 entries of SNPs associated with various cancers in the GWAS Catalogue (not all of these entries are unique) [73]. These SNPs are spread all over the genome. Most of the cancer associated SNPs are not located in protein coding regions and can consequently not affect the function of the protein [30, 45]. It is thought that many of these SNPs located outside of coding regions affect the regulation of transcription.

In this chapter two methods of linking SNPs to initial genes will be used. First the gene closest to the SNP will be considered to be the target gene. The second method

will be to consider the target genes of SNPs which act as Expression Quantitative Trait Loci (eQTLs). These are genetic variations associated with a differential expression of a transcript [74]. By considering eQTL target genes it is possible to take into account indirect interactions and identify more links that would be of interest. Furthermore, many of these eQTLs are not located anywhere near the target gene, and many times on different chromosomes [43] (*trans*-acting eQTLs). The presence of these *trans*-acting eQTLs point towards a limitation of the first method, using the gene closest to the cancer associated SNP in the study.

Once this small set of links has been found it is then much easier to search the literature for the most suitable model.

2.1.2 Identifying mathematical models of interest

Mathematical models often only handle one pathway, and often even a small part of a pathway. If one is to successfully link the effect of SNPs to the dynamics of mathematical models, it is therefore important to identify pathways, which both are enriched for relevant SNPs and also have been studied through high quality models. There are several ways of mapping genes to pathways and finding enrichments within them. Two of the most common pathway annotations are the Gene Ontology (GO) database [75, 76] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [77, 78, 79].

The GO database gathers information about genes regarding their function and classifies them in terms of biological process, cellular component (cellular location), and molecular function. Using the biological process annotations it is possible to link genes to pathways important in carcinogenesis, such as apoptosis, cell cycle or cell growth. While the KEGG database also contains classifications it has the very useful function of organising genes and proteins in pathways directly. It also contains disease related pathways and allows for focusing on cancer development for example. Although this might be very useful in many cases, if the goal is to identify models for further study, the results from enrichment analysis in these databases are too rich. Most genes will map to several pathways or biological processes and most of the mappings will not be covered by any available model.

Another approach to find models would be to directly map genes to nodes in models of interest and choose the models with the most target genes. Even though there are databases of dynamical models, such as CellML model repository [80], and BioModels Database [81], these do not in any way represent the vast amount of models that are available in literature. If one therefore were to map the SNPs onto the annotated genes in these databases one would in effect miss a lot of useful models.

Although Ordinary Differential Equation (ODE) models are usually limited to one pathway and are very small due to computational cost and limited knowledge about the details needed to parametrise the model, there are a number of Boolean models which are much larger in terms of nodes and cover larger parts of carcinogenesis. The Boolean model by Fumiã & Martins (2013) [53] is an example of such a model. It contains 96 nodes (compounds) and 259 edges (interactions between the nodes), capturing many of the important aspects of carcinogenesis. Due to its size and it spanning many cancer relevant pathways, this model was chosen as a starting point to identify links between breast cancer associated SNPs and relevant genes. From these genes, pathways will be identified, and promising models will be chosen for further study in the next chapters.

The reason for this model not being used to assess the effect of SNPs on the system dynamics (and for not considering any boolean models) is that it is difficult to model the effect of small changes in concentrations or activity in a boolean network. In a standard boolean network, a node can only be either on or off, meaning that if a SNP were to decrease the translation of a gene with 1% this would either have to be interpreted as the gene never reaching high enough concentration to be turned on, or that the change would not affect the dynamics of the model, since the node would be turned on either way. There are ways to model smaller changes in dynamics, such as introducing dummy nodes to model time delays or duplicating nodes to show increased activity. Stochasticity could also be introduced into the model so that each node has a given probability of turning on after downstream activation signal. A SNP could then have a small effect on the probability of the node to turn on. However, all of these methods have disadvantages as well, such as introducing arbitrary time scales, or requiring one to determine probability distributions for the activation of nodes. They also increase the complexity of the model and diminishes one of the big strengths of a boolean model, its simplicity. At some point it is simply better to use differential equation models instead, which are better suited for these kind of problems.

Although 96 nodes is a much larger number than in most ODE models, it is still a very small fraction of the total number of genes involved in carcinogenesis and it is not very likely that many direct links between cancer associated SNPs and these genes will be found.

To increase the number of interactions identified, Search Tool for the Retrieval of Interacting Genes/Proteins (STRING), a database of protein- protein interactions [82], will be used to identify intermediary interactions between the SNP associated gene and the model node gene.

2.2 Materials and Methods

2.2.1 Data sources

For the SNP analysis a data set of 85 curated breast cancer associated SNPs and their nearest genes collected from the Genome Wide Association Study (GWAS) Catalogue was used [83]. These had been filtered to only include SNPs with a p-value of lower than 10^{-5} (Table A.1). Often a p-value of 10^{-7} or even 10^{-8} is used as a threshold for a significant SNP. The less stringent criteria allowed for a higher chance of finding SNP-model connections, due to the increased sample size, albeit at the risk of introducing false positives. The strength of association between the genotype and the phenotype was reported in terms of odds-ratio:

$$\text{odds-ratio: } \frac{\text{Cancer(allele A)/Cancer(allele B)}}{\text{Healthy(allele A)/Healthy(allele B)}} \tag{2.1}$$

SNPs in linkage disequilibrium (LD) ($R^2 < 0.8$) with any of the breast cancer associated SNPs were extracted from the HaploReg database (version 2) [84]. Two SNPs in LD are associated with each other, i.e. their alleles occur together more often than would be expected by chance.

A set of eQTLs from blood samples were acquired from the Blood eQTL browser [85, 43]. This data lists, among others, the SNP rs-ID, the gene affected and a False Discovery Rate (FDR) score for the association between the SNP and the gene expression, where the significance of each eQTL had been tested against a null distribution of 10 repetitions of the analysis with permuted sample labels. eQTLs with a FDR < 0.05 were filtered for further analysis.

STRING (v9.1), a database containing protein-protein interaction data collected from a range of experimental and computational sources as well as literature research, was used [86]. Each interaction is scored based on the amount of evidence for the interaction. When data is available, the type and direction of the interaction is indicated as well. The data was filtered for interaction with a score of at least 800 which was suggested by the original paper to indicate strong support [82].

The nodes in the model by Fumiã & Martins are many times generic in the sense that they represent classes of proteins in the cell. The nodes were used to compile a list of proteins and different versions of the proteins covered by the model (Table A.2).

2.2.2 Linking SNPs to model genes

In order to find SNPs associated with breast cancer, which could affect the function of any model chosen for analysis, connections between the SNPs and the model genes were established according to the method outlined in Figure 2.1 and described below. As a starting point genes indicated to be targeted by eQTL acting, breast cancer associated SNPs in the whole blood data set by Westra *et al.* (2013) [43] were chosen. Proteins linked to these genes with a maximum distance of 2 interactions according to interaction data retrieved from STRING were then extracted [82].

Finally all 3-tuple protein sets with at least one node in the Boolean cancer model by Fumiã & Martins (2013) [53] were considered links between genotype and cancer phenotype. A distance of 2 interactions in this case would mean that there is support for an eQTL-affected protein A interacting with a protein B that in turn is interacting with a protein C that is present in the network. Those links which STRING indicated had a directionality going from an eQTL affected protein to a cancer model protein (that is, A is acting on B which is acting on C), were further extracted (Table 2.1).

The same procedure of extracting links of maximum length 2 was repeated for all the nearest genes associated with the original set of breast cancer associated SNPs (Table A.1). Additionally, the directed graph was examined, as described above.

Although the Boolean cancer model by Fumiã & Martins (2013) [53] is not being used directly in later stages of this work, it captures the core proteins of key pathways of cancer development and any model later chosen would likely target one of these pathways and consequently contain the protein modelled in this model as well as proteins directly interacting with it. Creating these maps of tuples allowed for identifying a larger set of genes which are likely to be within cancer related models.

Analysis was carried out in python 2.7.

Table 2.1: Genes targeted by eQTLs linked to breast cancer associated SNPs. For each target it is indicated whether the eQTL acts in a cis of a trans fashion.

eQTL target gene	eQTL type	eQTL target gene	eQTL type
ABHD3	trans	MAP3K11	cis
ANKLE1	cis	MARCH6	cis
ANKRD16	cis	MRPL34	cis
ATE1	cis	MTAP	cis
BANF1	cis	MUS81	cis
BBS7	trans	NR2F6	cis
C19orf60	cis	NSMCE4A	cis
C19orf62	cis	OCEL1	cis
C5orf35	cis	OR2A9P	cis
C6orf97	cis	PEX14	cis
CDKN2B	cis	PGPEP1	cis
CHMP4B	cis	PLAUR	cis
CTSW	cis	PLVAP	cis
DCLRE1B	cis	PNCK	trans
DFFA	cis	POP5	trans
ECHDC1	cis	PRRG4	trans
EFEMP2	cis	PTPN22	cis
EIF2S2	cis	RNF146	cis
ELL	cis	ROPN1L	cis
FAM89B	cis	SART1	cis
FGD5	trans	SIPA1	cis
FIBP	cis	SOCS1	trans
FPR1	cis	TEX9	trans
GPR68	cis	TGFBR2	cis
GTPBP3	cis	TMEM75	cis
HIPK1	cis	TNNT3	cis
KCNN4	cis	ZNF649	cis
KIAA1217	trans	ZNF613	cis
LRRC25	cis		

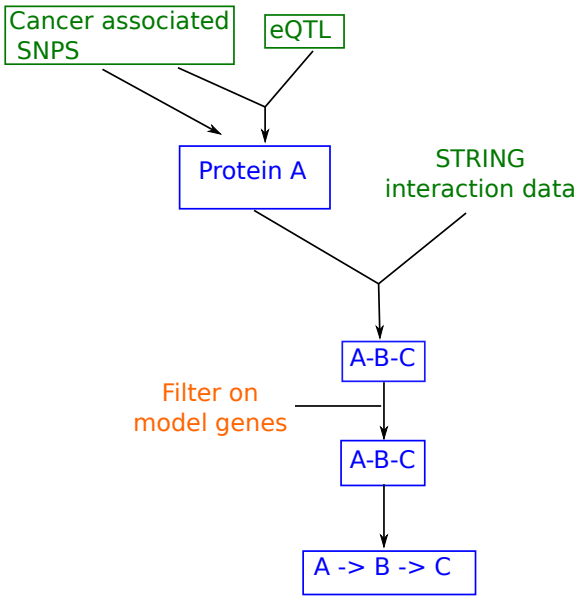


Figure 2.1: Flow chart of extracting 3-tuples of protein interactions. Proteins which interact with SNP or eQTL target genes according to STRING are extracted. Process is repeated once. 3-tuples where one protein is in the model by Fumiã & Martins are extracted. A further filtering is done, keeping those tuples where SNP or eQTL target gene is indicated to act on protein B, which is indicated to act on protein C.

2.3 Results

2.3.1 SNPs acting as eQTLs linked to 15 model genes

From the original 85 breast cancer associated SNPs 1966 SNPs could be retrieved from the HaploReg data base with an LD $R^2 > 0.8$. Out of these, 48 SNPs acted as *cis*-eQTLs and another 10 acted as *trans*-eQTLs in the blood eQTL data set. The *cis*-acting eQTLs (distance between SNP and midpoint of probe is less than 250 kb) could be linked to 50 proteins and the trans-acting eQTLs (distance larger than 5 Mb) to 9 proteins (Table 2.1). Using STRING, these proteins could be linked to a large number of proteins in the Boolean cancer network by Fumiã & Martins with a distance of 2 or lower (Figure 2.2). When only looking at the directed data, 6 cis eQTLs could be linked to 15 proteins in the model (Table 2.2 and A.3, Figure 2.3). In both cases, no eQTL affected genes were themselves a part of the model and in the directed graph, only two interactions were of length one (TGFBR2 - RTGFB1, TGFBR2 - SMAD7).

Table 2.2: Genes in the Boolean cancer model which are targeted either by a SNP associated with breast cancer or by an eQTL linked to any of the previously mentioned SNPs.

Model Gene	Whole SNP Data	eQTL SNP
ATM	1	-
CCNB1	1	1
CCND1	1	1
CCNE1	1	1
CCNE2	1	1
CDC20	1	1
CDKN1A	1	1
E2F1	1	-
E2F2	1	-
E2F3	1	-
E2F4	1	-
E2F5	1	-
EEF2	1	1
HIF1A	1	1
IKBKB	1	-
MAP3K7	1	-
NFKB1	1	-
NFKB2	1	-
RB1	1	-
SF3B6	-	1
SMAD4	1	1
SMAD7	1	1
SMAD9	-	1
TERT	1	-
TGFB1	1	1
TNF	1	-
UBE2C	1	1
VHL	1	1

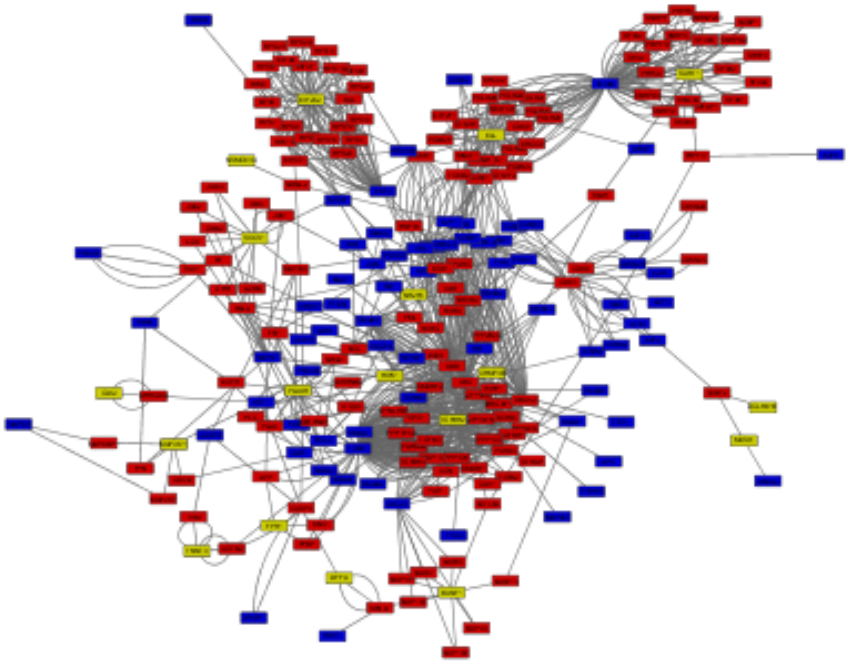


Figure 2.2: Network of breast cancer associated SNPs acting as eQTLs interacting with cancer model genes. Proteins marked in blue were identified as eQTL genes from the 1966 breast cancer associated and LD SNPs. Each protein interacts with a protein in the Boolean cancer model by Fumiã & Martins, either directly or with one intermediary interaction. Red nodes are genes in the model and yellow nodes are intermediary interactions.

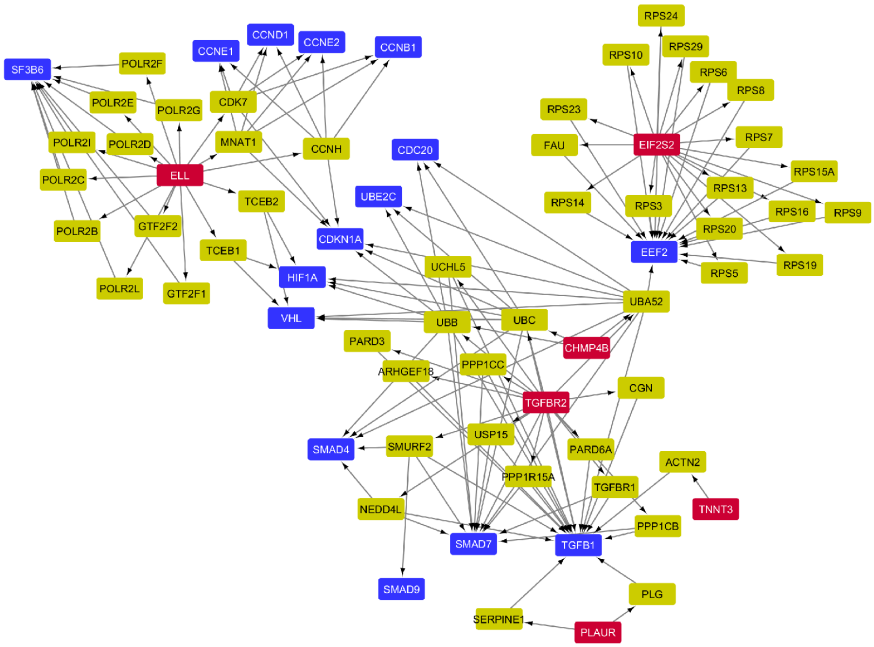


Figure 2.3: Subset of network in Figure 2.2, where all interactions have the right directionality as indicated by STRING. Red nodes are eQTLs interacting with cancer model genes nodes (blue), either directly or through an intermediary interaction (yellow).

2.3.2 Linking SNPs to nearest gene resulted in more connections

Directly linking the SNPs from the GWA studies, instead of going through eQTLs, also resulted in a large and very complex network of interactions of length < 3 (Figure 2.4). Considering only interactions which indicated that the nodes in the cancer network are being acted upon and the SNP associated proteins are the initial actors resulted in a network of 6 proteins associated with GWAS SNPs acting upon 26 nodes in the network. Once again, most interactions are of distance two, but 9 are of distance one and 2 (TERT and CCND1) are a part of the network themselves (Figure 2.5 and Table A.4).

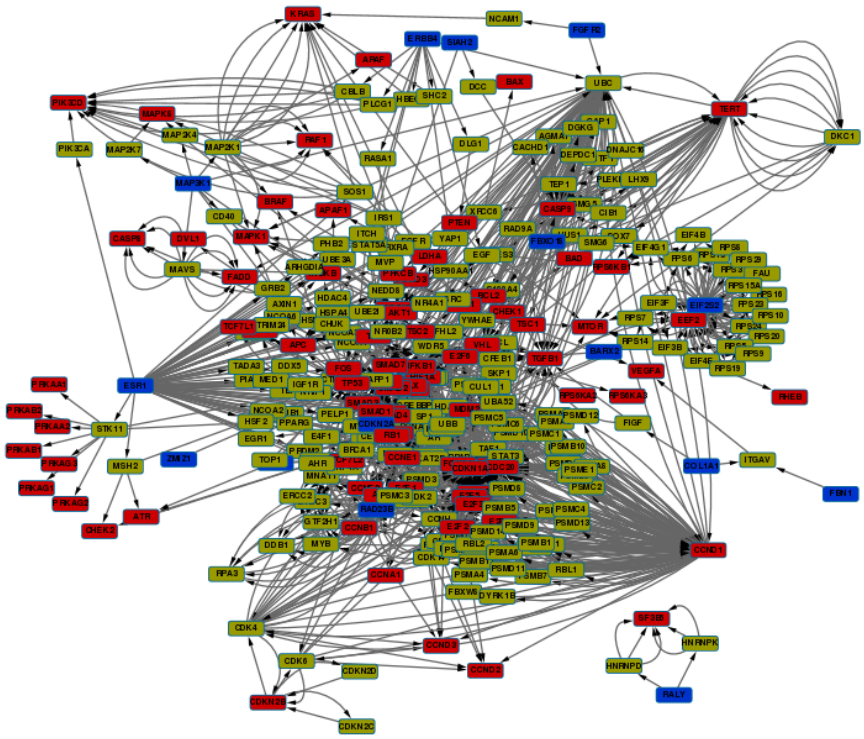


Figure 2.4: Network of interactions between genes mapping to breast cancer associated SNPs (blue) and genes in the Boolean cancer model by Fumiã & Martins (red). Yellow nodes are intermediary interactions. Most interactions are going from model nodes to SNP genes, but a large portion of them are also going in the relevant direction, that is SNP associated genes are affecting model genes.

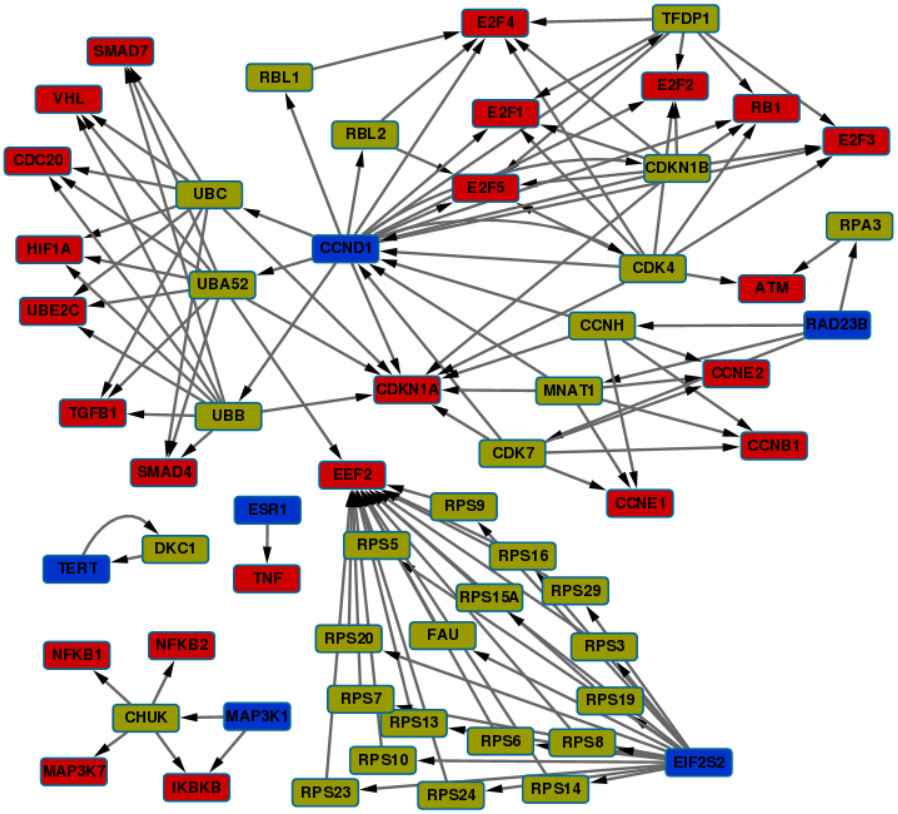


Figure 2.5: Network of interactions between breast cancer associated SNPs and nodes in the Boolean cancer model by Fumiã & Martins. All interactions go from SNP gene (blue) to model gene (red), either directly or through an intermediary interaction (yellow). In addition, two of the SNP genes, CCND1 and TERT, are themselves part of the model.

2.4 Discussion

2.4.1 Linking SNPs to genes and pathways

Although the breast cancer associated SNPs expanded to 1966 SNPs when including all LD SNPs, it quickly reduced back to a manageable amount of genes when only looking at eQTL SNPs. The same trend could be seen when first mapping the genes to the cancer network model through interactions covered in STRING and then only considering the interactions with the right directionality.

2.4.2 Identifying mathematical models of interest

By investigating the directed graphs of interactions, in Figure 2.3 and Figure 2.5, a number of possible networks emerged for further studies. One promising pathway, which emerged both when studying the graph of breast cancer associated SNPs and the eQTL SNPs was the cell cycle. This network involves CCND1 (Cyclin D, mapped directly to a breast cancer associated SNP), and the other cyclins CCNB, CCNE1 and CCNE2. Other important genes in this pathway are CDKN1A, RB1 and all the E2F genes, also implied to be affected by breast cancer associated SNPs. The cell cycle has been studied for a long time and there are many interesting models of various complexities for studying this system [55, 87, 56, 88, 58, 57].

Other interesting interactions were also found, such as the one involving ATM, a gene involved in the DNA damage response [28, 89], and are very likely to be important for carcinogenesis. This gene also has a big effect on cell cycle progression [90]. However, at the time only Boolean models of the DNA damage response pathway could be identified. Considering the small effect eQTLs and most SNPs have on a gene, the binary states used in such a model would not be able to represent such changes using standard analysis tools (as explained in Section 2.1.2). Once ODE-models of these pathways have been developed, this will be a very interesting path to follow.

The cell cycle was the most enriched pathway, probably due partly to it being an essential pathway, but also because the Boolean model itself is enriched in cell cycle related proteins. Initially a model of the cell cycle was chosen for further study [58]. However, the phenotype of the model proved to be very robust with regards to parameter perturbations (data not shown). Upon perturbation of any parameter, the period of the cell cycle shifted as expected. However, when the model was let to run for several cycles, it slowly reverted back to the initial period. This was true for large ranges of parameter perturbations and only in some cases did the model shift towards a cell cycle of a different length, but then the period of that cell cycle showed a similar behaviour. This is likely

to be a feature of the model selected and not of all cell cycle models. However, because the cause of this behaviour could not be understood and the effects it would have on any future analysis was unclear, it was abandoned and focus was concentrated on the apoptosis models which also had strong links to the SNP data.

As a test case for linking SNPs to dynamical models the apoptosis model by Schlatter *et al.* (2011) [91] was chosen. It is a good example of how one can combine several smaller models into one and contains several links to the SNP data. The eQTL gene MAP3K11 is known to work in TNF mediated activation of JNK [92] and the SNP associated gene MAP3K1 is suggested by the interaction data to act upon several proteins in the NFκB pathway. The SNP associated gene ESR1 is also suggested to affect the input node TNF. In addition ESR1, an eQTL target of the extended set of SNPs is indicated to be a transcriptional regulator of several genes in the apoptosis signalling pathway [93]. Likewise NR2F6 is shown to regulate the expression of X-linked inhibitor of apoptosis protein (XIAP) and possibly the two apoptosis regulating genes Bax and Bcl2 [94]. In addition to the large apoptosis model by Schlatter *et al.*, a smaller apoptosis model by Eissing *et al.* (2004) [95] was chosen. Several SNPs in both the breast cancer and prostate cancer data set used in Chapter 5 map to genes suggested to have binding sites in promoters of genes in this model.

The smaller apoptosis model covers the core of the apoptotic signalling pathway, involving Caspase 3 and Caspase 8 as well as the two inhibitors IAP and BAR. The simplicity of the model renders it suitable for exploring the different methodologies in detail and will prove very useful as the project develops in the later chapters. The larger model also covers the core around the Caspases, but in addition also covers a much larger part of upstream proteins, all the way up to the two membrane bound receptors TNFα and Fas. It also models reactive oxygen species and includes an NF-κB module to model the effect of TNFα activation on the transcription machinery.

In the following chapters both models will be used to explore different sensitivity analysis methods and how the results from them can be used to understand the risk of acquiring a cancer phenotype. Then a new phenotype sensitivity analysis tool will be developed and both models will be used to theoretically explore the effect of the genotype on the risk of developing a cancer phenotype. Finally, the smaller model will be used to validate the theoretical results on experimental data for breast and prostate cancer.

Chapter 3

Sensitivity Analysis

3.1 Theory

In chapter 1 six challenges were outlined, which had to be solved during this thesis. In chapter 2 the first 2 challenges were addressed whereby first, Single Nucleotide Polymorphisms (SNPs) known to be associated with the risk of developing breast cancer were linked to genes and pathways involved in carcinogenesis and second, a smaller and a larger apoptosis model were identified, which could be linked to breast cancer associated SNPs.

In this chapter the third and fourth challenges will be addressed. Various features of the models will be assessed for their suitability to represent the two phenotypes corresponding to normal and abnormal response to apoptotic signalling. Additionally, methods to assess the sensitivity of the model features to perturbations in the models corresponding to the presence of a SNP will be explored. These two challenges will be addressed at the same time, as the features available for assessment will be determined by the type of analysis method used.

Although only the two apoptosis models will be used in this chapter, the hope is that the analysis performed in this thesis will be applicable to models of any hallmark of carcinogenesis. Therefore, all principles and theoretical framework needed will first be discussed from a general point of view, before examples will be made of how these principles apply to apoptosis and specifically the two models chosen.

3.1.1 Model behaviour

When studying dynamical system models there are a range of tools and concepts which are generally used. The most basic of these concepts is that of variable trajectories. Studying these trajectories provides information about the changes in variable values over time. By

studying the trajectories of a system, the behaviour of the system as a whole, given a set of initial conditions (ICs), can be deduced.

The behaviour of most deterministic, dynamical systems can roughly be divided into four groups (Figure 3.1). In the first scenario the values of the system's variables could continue to grow indefinitely like a population not being limited by space, nutrient availability or growth suppressors. In the second scenario the dynamics of the system could eventually settle on some fixed behaviour: The system is said to have reached an attractor. This type of behaviour can be seen in the way a population limited by space and resources eventually reaches a steady state (SS) where its growth rate is the same as its decay rate, and the population density is constant over time. It can also be seen in the way molecular signalling within a cell upon activation increases and later maintains a constant level of signalling. In these cases the attractor is a fixed point of concentrations or amounts of all components. A third group of behaviour is when the values of the system variables repeat themselves in a predictable way. An example of this could be how the levels of proteins involved in the cell cycle oscillate in a time dependent manner. Here the attractor is not a fixed point but each component follows a chain of states which eventually closes on itself. A fourth type of behaviour, which will not be covered in this chapter, is that of a chaotic system. In these cases the components increase and decrease in a way which never repeats itself.

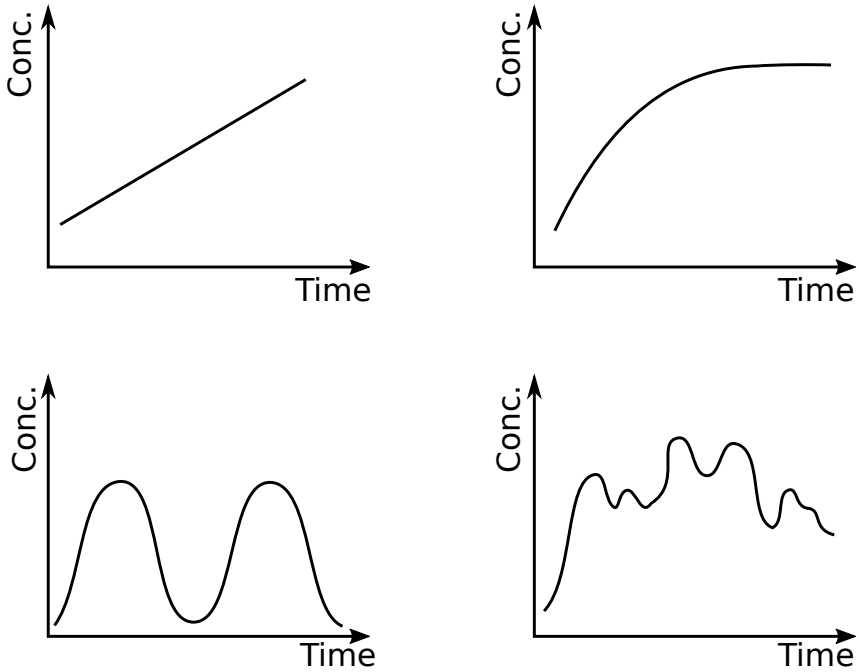


Figure 3.1: Illustrations of the main types of dynamics in dynamical systems. (A): Continuous growth without any upper limit. (b): The dynamics eventually reaches a steady state where it remains. (c): The dynamics fallow a repeating pattern. (d): A completely chaotic, non-repeating pattern.

Depending on the ICs of a system, it could show different behaviours and it is possible for a system to have several competing attractors. For example, consider a system with a competing activating and deactivating function. Depending on which of the two functions is stronger, the system will either move towards a state of activation or deactivation. In simpler systems, by drawing these functions in the same diagram, it is possible to predict in which direction the system will move given any combination of its variable values (Figure 3.2). This information can be summarised in a phase space, where all variables of the system are represented on one axis each. At any point in this space, it can be calculated, in which direction the state of the system will be moving, and by connecting these trajectories it can be determined in which, if any, of the attractors, it will end up (Figure 3.3). For example, it is possible to establish whether an activation signal of a certain strength will propagate through a system and turn it on, or if it will die out due to the intrinsic inhibitory functions. These two attractors will each be represented by a single point in the phase space, whereas an attractor like that of the cell cycle will be a a closed loop. Any points in phase space where the system is drawn towards a specific attractor is called the basin of attraction. The border between these basins are called

separatrices, since they separate different types of dynamics. Both of these features of the phase space are very important when linking the model dynamics back to the biology and the risk of developing cancer, as they define the potential of the system to change phenotype.

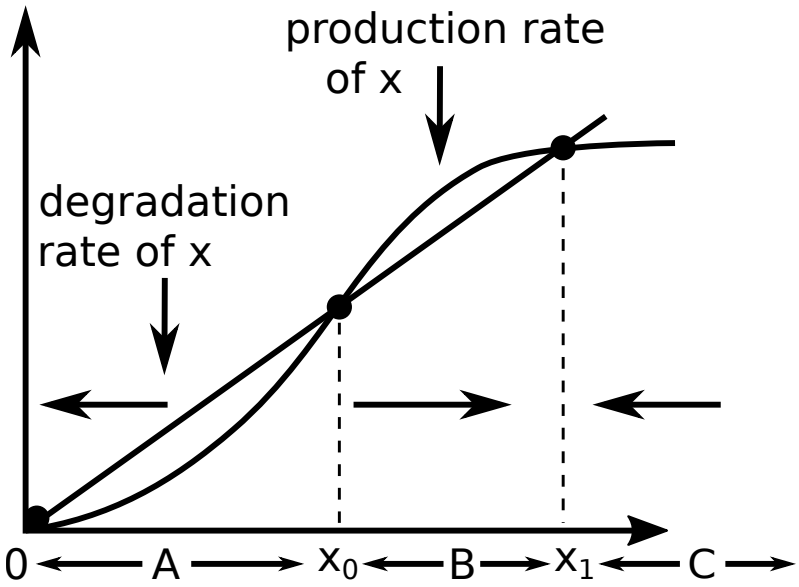


Figure 3.2: Illustration of rates of production and rate of degradation for a two-equation system. Given the two curves it can be established for any value of x, if the concentration will increase or decrease. If x is such that the two curves cross, the system is in equilibrium and will not move. If it is between the first and the second intersection, or after the third intersection (A and C), the degradation rate is higher than the production rate and x will decrease until it reaches an equilibrium. If x is between the second and the third intersection (B), x will increase until it reaches the third intersection.

The tools of dynamical systems analysis can also be used to gain an understanding of how parameter changes can affect the system. Given the equations of the system it can be deduced how a change in a parameter alters the positions of the attractors in the phase space. Given certain parameter changes, it is also possible to alter the system so much that a bifurcation occurs, where some attractors disappear or others appear (Figure 3.4). This has huge implications when considering the biology behind the modelling. Such shifts in the dynamics could not only affect the strength of a signal, or whether it would be activated given certain starting conditions, it could also prove the system to be unable to reach such a SS at all.

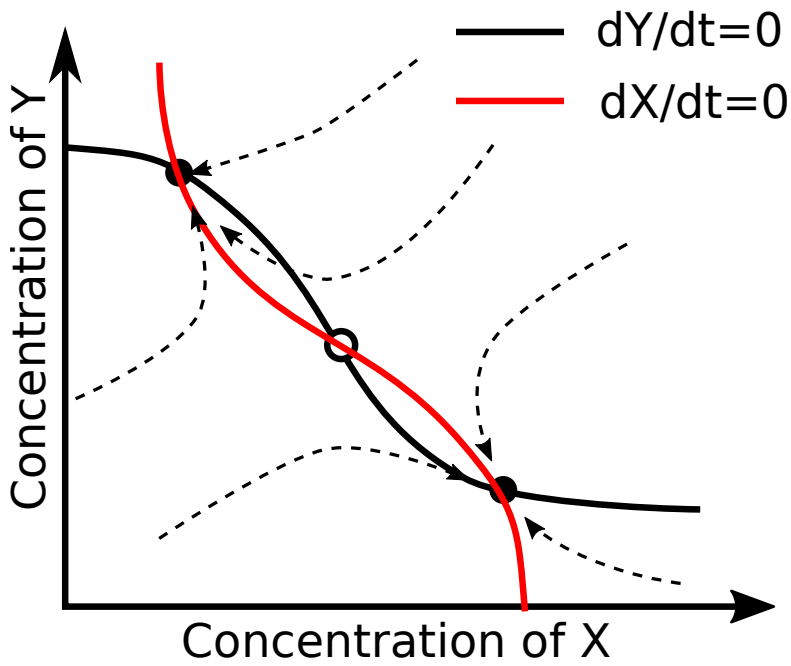


Figure 3.3: Phase portrait of a two-variable system with two attractors and one unstable steady state. If the system starts exactly in the middle steady state it will remain there. If it however, initiates anywhere else, it will fall into one of the other attractors. For any point in the portrait, it is possible to know where the system will move next and consequently the path from any point to its final attractor state can be drawn (arrows).

When applying these methods to the genetics of cancer development pathways, the normal state of a system, starting with a given set of parameters and variables, could be considered as occupying a small volume of the phase space, in which the system will move towards the attractor corresponding to normal behaviour of the system. For example, when considering apoptotic signalling, the default conditions would result in the system reaching a SS corresponding to the onset of apoptosis, given an initial activation signal. Depending on the model at hand, it is possible to then consider mutations which would cause either a shift of a variable, causing the system to start in a different position of the phase space, or a shift in a parameter, causing the SSs in the phase space to change in location or number (as explained in Figure 3.4). It is of course also possible to consider models where both variables and parameters could be altered by mutations. Either type of alteration could result in apoptosis not occurring.

Even if the phase space contains the information on where the system will end up depending on any starting point, it does not directly contain information about how

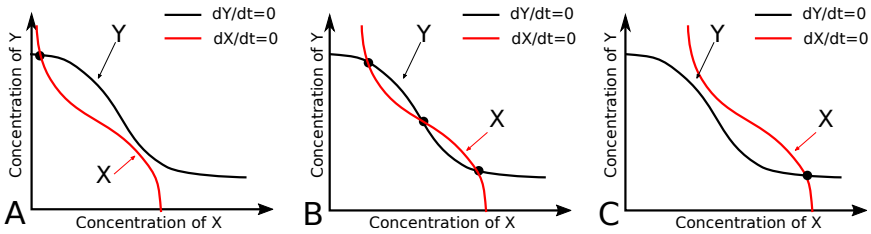


Figure 3.4: Illustration of how steady states emerge and disappear as parameters are perturbed. A: Self activation of X is weak, resulting in only one stable steady state with high concentration of Y and low concentration of X. B: as the self-activation of X gets stronger a second stable steady states and an unstable steady state in-between the two emerges. C: as the self-activation gets even stronger the first steady state disappears.

long it will take to reach that attractor. When considering biological processes, this is important, since a system might have attractors, which may be mathematically valid, but would take so long to reach that they are not biologically relevant. For example, it might be possible for an apoptosis signalling model to reach an attractor corresponding to onset of apoptosis, but it might take days, weeks or even years to do so. It is also important to keep timing in mind in work like that performed in this thesis, where the models will not be analysed analytically, but simulated within a time frame deemed relevant. This means that any mutation does not have to cause a disease associated SS to disappear. It just has to move the attractor far away enough for it to be non-approachable within a relevant time frame.

To sum up, the dynamical systems analysis involve the study of trajectories of variables in the system, the occurrence of any SSs, positions of attractors in phase space and the location of basins of attraction as well as the locations of the borders between the basins (separatrices), given a set of ICs and parameters. Changes in parameter values can cause changes in all of the above mentioned characteristics of a system and may result in changes in behaviour of the system. Some of these changes may be qualitative, for example, gain or loss of attractors (bifurcation). They may involve changes in behaviour due to ICs moving between basins of attraction. They may also change the timing of the dynamics, which may be crucial in a biological system.

From studies of the effects of SNPs it is known that they often do not have very large effects on the system under normal conditions [96, 97]. For example, a SNP altering the expression of a gene important in apoptosis signalling might not have a measurable effect on the cell's ability to undergo apoptosis under normal conditions. This is not surprising, since it could be considered a key evolutionary advantage to have robust behaviour of core functions such as apoptosis and these systems would then have evolved such that

the normal state is positioned deep within the basin of the desired attractor. However, if additional mutations alter the shape of the phase portrait, or the starting position within the portrait, it is possible that the small shift caused by the initial SNP could significantly alter the system dynamics and push it out of the basin of attraction, i.e. the part of phase space where the system will be attracted towards that specific attractor.

3.1.2 Sensitivity

A common concept within systems biology and a red thread throughout this thesis will be the concept of sensitivity. Sensitivity is a very important topic in systems biology and there are numerous ways of assessing the sensitivity of a system. However, there is no universal definition of sensitivity and depending on the underlying question, sensitivity can be defined in many different ways, and consequently the ways in which it is being assessed, and the conclusions which can be drawn from the analysis can vary significantly. It is therefore paramount to clarify the various ways this term will be used in this thesis, and how they relate to each other and the overall question.

One notion of sensitivity which will be explored is that of model output with respect to perturbations in ICs, i.e. how much does the output of the model change, when the ICs change. In terms of the phase space, it asks how much does the path to the attractor change when the system initiates in a different position. If the ICs are deep within the basin of an attractor any changes, while still changing the exact trajectory of the system, are unlikely to cause the system to end up at a different attractor. It is therefore said to have low sensitivity to IC perturbations. If the ICs on the other hand are close to the separatrix, the system may end up moving towards a different attractor if the perturbation is given in the right direction. Such a system would have a high sensitivity to ICs (Figure 3.5).

Another type of sensitivity is called parameter sensitivity. This refers to the change of the system dynamics upon perturbations of its underlying parameters, i.e. how do the attractors and the trajectories of the phase portrait move upon parameter perturbations. If the ICs are deep within the old basin of attraction, any perturbation, while moving the exact position of the attractor, is unlikely to move the system out of that basin. If the ICs on the other hand are close to the separatrix of the original phase space, a perturbation may cause the separatrix to shift so that the ICs end up in a different basin of attraction (Figure 3.6).

For most of this thesis, the detailed dynamics of the system are of minor concern. What is really of interest is the final behaviour of the system and its sensitivity to genomic and meta genomic perturbations. In the example of apoptosis, it is not the detailed dynamics

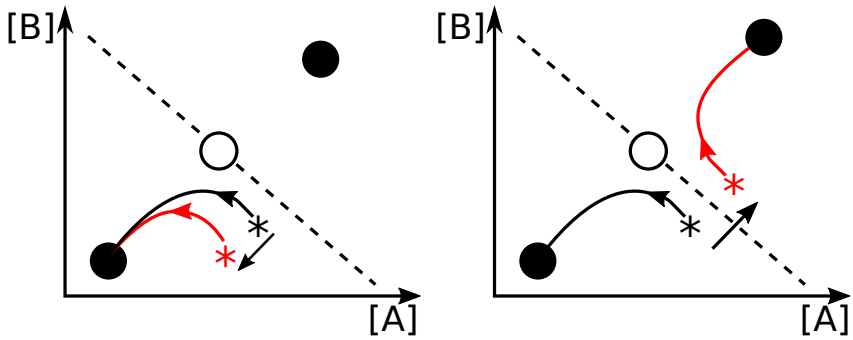


Figure 3.5: Sensitivity to ICs. Left: A shift in ICs causes the trajectory of the system to change. However, it is still drawn towards the same attractor as the original ICs. Right: a shift in IC causes the system to cross the separatrix and to be drawn towards a different attractor.

of Caspase concentrations, or the rate of phosphorylations and how much they differ with or without a given mutation over the course of the cell signalling process, which matters. What is really of interest, is how much these effects change the parts of the dynamics which can be interpreted as the time to apoptosis. So it is not the difference in the path in phase space which matters, but rather, whether the system is pulled towards the attractor corresponding to apoptosis and how long it takes to get there. However, since crossing into a different basin of attraction will have a major effect of the trajectories, it may still be possible to study this sensitivity by studying the change in time course behaviour.

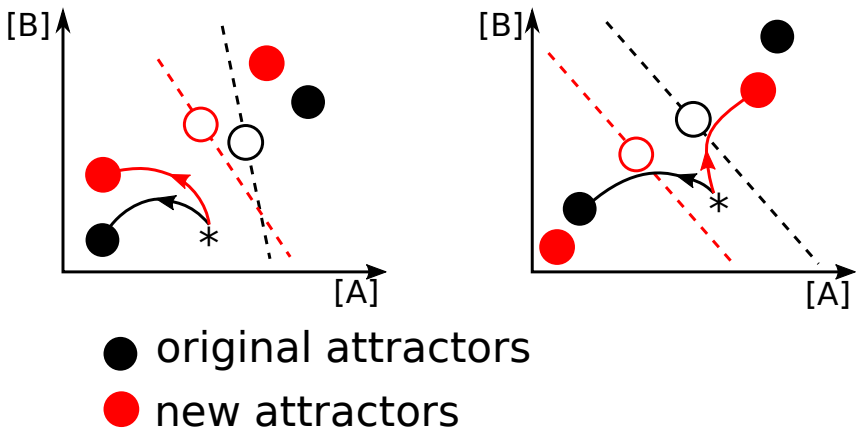


Figure 3.6: Sensitivity to parameter changes. Left: a change in parameter values causes the attractors to move in phase space. However, the ICs are still within the same basin of attraction as it was before and moves towards the new lower attractor. Right: a change in parameter values causes the attractors to change so much that the ICs end up on the opposite side of the new separatrix and is therefore drawn towards the upper attractor.

3.1.3 Sensitivity analysis tools

Generally the types of sensitivity analysis are divided into local and global sensitivity analysis. In their simplest forms, both types of methods treat parameters as independent random variables. This is usually a reasonable assumption to make in biological systems as for example the production rate of one protein would not affect the production rate of any other protein during normal conditions due to the excess capacity of the transcriptional and translational machinery in the cell. Local sensitivity considers the change in dynamics at each time point after perturbation of a single parameter. One of the simplest versions make use of the fact that the change in output at any given point can be written as a Taylor series around that point:

$$y_i(t, k + \Delta k) = y_i(t, k) + \sum_{j=1}^{m} \frac{\partial y_i}{\partial k_j} \Delta k_j + \frac{1}{2} \sum_{l=1}^{n} \sum_{j=1}^{m} \frac{\partial y_i}{\partial k_l \partial k_j} \Delta k_l \Delta k_j + ... \tag{3.1}$$

where k is a vector of m parameters.

The second term in the Taylor series is then the first order sensitivity. This can be approximated by the difference between the original point and the point after introducing the perturbation normalised by the perturbation:

$$s_{i,j} \approx \frac{y_i(t, k_j + \Delta k_j) - y_i(t, k_j)}{\Delta k_j} \tag{3.2}$$

By introducing higher order sensitivity terms co-sensitivities between parameters can be considered. However, the local sensitivity has to be calculated for each perturbation or set of perturbations considered.

Global sensitivity on the other hand tries to handle the sensitivity in a more general fashion. Often a Monte Carlo sampling is applied to a range of valid parameter values and the sensitivity is calculated as a function of the change in output over that range.

One common global sensitivity analysis method is the Morris method [98]. This is very similar to the local sensitivity analysis method just mentioned, but is performed several times to cover a part of parameter space as large as possible. The mean and spread of the sensitivities gathered during these iterations can then be used to study the global sensitivity.

If the sensitivities are linear, a linear regression can also be performed and the sensitivity can be analysed in terms of correlations between outputs at various perturbations.

Another common global sensitivity method is Sobol's method [99, 100]. This is a variance-based method and calculates the contribution the variance in each parameter has on the total variance of the output. As long as the parameters are independent they

can come from any probability distribution. However before using the method the space from which the parameter values are sampled has to be transformed into a standard uniform distribution ($X_i \in [0,1]$). If this is done, then given Equation 3.1 and the fact that $Var(A+B) = Var(A)+Var(B)$, the total variance of the output can be written as:

$$Var(y_i(t,k+\Delta k)) = \sum_{j=1}^{m} Var(\frac{\partial y_i}{\partial k_j}\Delta k_j) + \frac{1}{2}\sum_{l=1}^{n}\sum_{j=1}^{m} Var(\frac{\partial^2 y_i}{\partial k_l \partial k_j}\Delta k_l \Delta k_j) + ... \quad (3.3)$$

where Δk_j is the difference between the sampled parameter value and the original value.

The first order Sobol sensitivity index can then be written as:

$$S_j^i = \frac{Var(\frac{\partial y_i}{\partial k_j}\Delta k_j)}{Var(y_i(t,k+\Delta k))} \quad (3.4)$$

By considering higher order interactions, covariances can also be studied. There are also adaptation to Sobol's method which allows for analysis of dependent variables.

3.1.3.1 Sensitivity analysis tools used in this thesis

In this thesis, two previously published sensitivity analysis tools will be used: SASSy [101] and SloppyCell [102]. Both methods look at the sensitivity locally in the sense that they only take into account a small perturbation from the original parameter values, as opposed the entire range of parameter values of interest. However, they do that for each time point over the entire time course of interest and for all parameters. It is therefore an important assumption that the change in dynamics when crossing the separatrix will be much larger than any changes due to a small shift of the original attractor. Although both methods try to assess the sensitivity of the system output, they do so in slightly different ways and it is important to understand how the differences in methodology and aim affect the results, why they differ from each other and how this relates to the questions addressed in this thesis.

3.1.3.1.1 SASSy

SASSy is a toolbox for sensitivity analysis of dynamical systems. The main strengths of this toolbox are the two graphical tools: Sensitivity Heat Map (SHM) and Parameter Sensitivity Spectrum (PSS). For the purpose of this study only PSS will be used.

In short, SASSy considers a differential equation $dx/dt = f(t,x,k)$ where k is a set of parameters. The solution to such an equation $x(t) = g(t,k)$ with the initial condition

$x(0) = g(0, k)$. It then calculates the effect of perturbations of all parameters by considering a discretised time frame $t = t_1, ..., t_N$ and creates a vector r for each state variable x_m $(m = 1, ..., n)$ and parameter k_j such that $r_{j,m} = (\partial g_m(t_1, k)/\partial k_j), ..., \partial g_m(t_N, k)/\partial k_j)$. The derivatives are calculated analytically for the output of the model at the discrete time points. For each j, these vectors are concatenated into a vector r_j which then forms the jth column of the matrix M such that:

$$M = \begin{bmatrix} \frac{\partial g_1(t_1,k)}{\partial k_1} & \frac{\partial g_1(t_1,k)}{\partial k_2} & \cdots & \frac{\partial g_1(t_1,k)}{\partial k_s} \\ \frac{\partial g_1(t_2,k)}{\partial k_1} & & & \\ \vdots & & & \vdots \\ \frac{\partial g_1(t_N,k)}{\partial k_1} & & & \\ \frac{\partial g_2(t_1,k)}{\partial k_1} & & & \\ \vdots & & & \vdots \\ \frac{\partial g_n(t_N,k)}{\partial k_1} & & & \frac{\partial g_n(t_N,k)}{\partial k_s} \end{bmatrix} \tag{3.5}$$

It then uses singular value decomposition (SVD) of M, $M = UDV^t$ to calculate the sensitivities of the system with respect to each parameter. In the SVD equation D is a diagonal matrix of singular values $\sigma_1, ..., \sigma_s$ of M, U is an $nN \times s$ orthonormal matrix and V is $s \times s$ orthonormal matrix. From this the PSS can be calculated as $S_{i,j} = \sigma_i V^t_{i,j}$ (note that this $S_{i,j}$ is unrelated to S^i_j used in equation 3.4).

3.1.3.1.2 SloppyCell

SloppyCell is a multi purpose tool which, among other things, can both fit parameters of a model to experimental data and calculate the sensitivity of the system with regards to parameter perturbations.

The sensitivity analysis is performed by calculating the average squared change in variable time course as parameters θ are perturbed from the reference values θ^*:

$$\chi^2(\theta) \equiv \frac{1}{2N_c N_s} \sum_{s,c} \frac{1}{T^c} \int_0^{T^c} \left(\frac{y_{s,c}(\theta, t) - y_{s,c}(\theta^*, t)}{\sigma_s} \right)^2 dt \tag{3.6}$$

where N_c is the number of conditions under consideration (in this thesis there is only one condition considered each time) and N_s is the number of species in the system.

Analysing the Hessian:

$$H^{\chi^2}_{j,k} \equiv \frac{d^2\chi^2}{d\log\theta_j d\log\theta_k}, \tag{3.7}$$

at θ^* then corresponds to approximating the surface of constant model behaviour change as N-dimensional ellipsoids where N is the number of parameters (note that in this case first order changes are zero since θ is evaluated at the "optimal" value). The reason for considering $\log\theta$ is that parameters can vary widely in scale.

If $H^{\chi^2}_{j,k}$ is evaluated at θ^* it can be calculated as:

$$H^{\chi^2}_{j,k} = \frac{1}{2N_cN_s}\sum_{s,c}\frac{1}{T^c\sigma_s^2}\int_0^{T^c}\frac{dy_{s,c}(\theta^*,t)}{d\log\theta_j}\frac{dy_{s,c}(\theta^*,t)}{d\log\theta_k}dt. \tag{3.8}$$

Using the Hessian, a Principal Component Analysis (PCA) can be performed where the eigenvectors end up being aligned with the ellipsoids of constant model behaviour change and the width of the ellipsoids are proportional to $1/\sqrt{\lambda}$ where λ is the corresponding eigenvalue

It is worth pointing out that the method analyses the sensitivity in trajectory change of the species, and not the sensitivity in system outcome. This means that extra care will have to be taken when interpreting the results and how it relates to the time of apoptosis.

3.1.4 Models

Two apoptosis signalling models of different degrees of complexity and slightly different characteristics were chosen as test cases; a TNFα and Fas induced apoptosis signalling model published by Schlatter *et al.*, (2011) [91] and a simpler model by Eissing *et al.* (2004) [95] which captures the core of the apoptosis signalling pathway. These are referred to as the larger and the smaller apoptosis model, respectively.

The smaller model captures the core behaviour of the apoptotic network. It consists of 4 proteins: Caspase 3, Caspase 8, BAR and IAP. The idea is that an upstream pathway activates Caspase 8, which in turn activates Caspase 3. Caspase 3 also activates more Caspase 8 to form a positive feedback loop and an increased signal strength. Activated Caspase 3 then triggers apoptosis through processes not covered in the model. BAR and IAP on the other hand act as inhibitors for Caspase 8 and Caspase 3 respectively. They achieve this partly by binding the active form of their respective target and thereby preventing it from propagating the activation, but also by targeting the complex for degradation, thereby depleting the pool of Caspase 8 or Caspase 3. All four proteins are

continuously being degraded and produced to maintain a SS of no Caspase 3 activation without the presence of activated Caspase 8.

Even though Caspase 3 is continuously being activated by the active form of Caspase 8, and the active form of Caspase 3 in return activates more Caspase 8, both active forms are rapidly being taken up by their respective inhibitors and consumed. However, once the inhibitors have been depleted by the continuous activation of Caspase 3 and 8, there is a rapid burst of Caspase 3 and 8 activation, which quickly reaches a new steady-state (Figure 3.7).

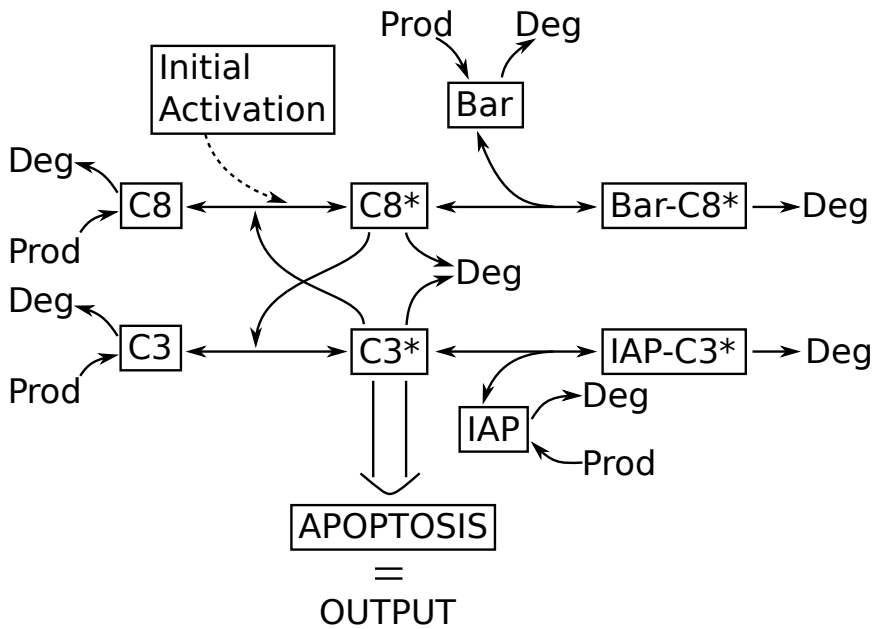


Figure 3.7: Network of small apoptosis model. The model takes an initial amount of activated Caspase 8 as input. The activated form of Caspase 8 activates Caspase 3, which in turn activates more Caspase 8. Both the active form of Caspase 8 and Caspase 3 are quickly being bound by BAR and IAP, respectively, and the two complexes are slowly being degraded. Both complexes can dissociate into their initial components and the concentrations of the four proteins Caspase 3, Caspase 8, BAR and IAP are being governed by constant production and concentration dependent degradation. If the activation of the two Caspases is faster than the production of the two inhibitors and the degradation of the complexes, it will eventually result in a burst of active Caspase 3, which can be interpreted as a commitment to apoptosis.

On a behavioural level, this system is bistable, very much like the example in Figure 3.3, with two stable SSs, one where the concentrations of activated Caspase 8 and 3 are close to zero (the initial conditions) and one where they are higher (the potential onset of apoptosis), separated by an unstable SS. The fact that the concentrations of all proteins

are governed by production and degradation parameters, means that there is only one set of biologically possible ICs, namely the lower SS, as the higher SS would mean the cell has committed to apoptosis. Before any signal from within the cell or from the surrounding, the system would start off, without activated Caspase 8, at the lower SS. The introduction of activated Caspase 8 would then push it across the separatrix into the basin of the higher attractor, i.e. committing to apoptosis. A perturbation of the initial activation of Caspase 8 may render it insufficient to push the system over to the basin of the higher attractor and it will revert to the lower attractor.

If the activation signal is held constant, a parameter perturbation, causing the unstable SS to move closer to the higher SS such as the example in Figure 3.4, can result in a previously sufficient amount of Caspase 8 activation, not being able to move the system across the unstable SS and the system would revert to the lower attractor.

The fact that the ICs, except for activated Caspase 8, are governed by parameters makes it very suitable for the two parameter sensitivity methods used in this chapter. This means that the sensitivity methods could evaluate the effect of SNPs which would cause a change in either activity or expression level of a protein. The analysis of a perturbation of a parameter affecting the function of a protein would be straightforward as such a perturbation would only affect the shape of the phase space and not the IC. A perturbation of a parameter affecting protein expression on the other hand would also shift the location of the lower SS, which is also serving as IC. For very small perturbations, it can be assumed that the new SS would be reached almost immediately and therefore the difference between the old and the new IC would not significantly affect the trajectory of the system in phase space. However, if the perturbations are large the different ICs could result in very different paths in phase space. To avoid this, whenever such perturbations are introduced in this analysis, the system will first be run to its new SS (without any activated Caspase 8) and the variable values at this SS will be used as new IC.

The larger apoptosis model captures a much larger part of the apoptosis signalling pathway (Figure 3.8). The larger size means there is a higher probability for SNPs to target the system, which makes it more likely to be able to link any identified sensitivities to biological data of disease associated SNPs. Like the smaller apoptosis model, the larger model captures the core of the apoptosis signalling network, where Caspase 8 activates Caspase 3 which then triggers apoptosis through mechanisms not covered by the model. Like in the smaller model, Caspase 3 is bound up and degraded by an inhibitor, in this case XIAP. Caspase 8 is also being degraded in a concentration dependent manner. In addition to directly activating Caspase 3, Caspase 8 also activates Bid, which activates BaxBak, causing cytochrome C to be released, which in turn also activates Caspase 3.

The BaxBak-cytochrome C mediated activation of Caspase 3 can also be activated by Bim, which in turn is activated by an upstream pathway. Whereas the smaller model takes an initial activation of Caspase 8 as input, in the larger model this activation is modelled by a Fas signal which activates Caspase 8 through DISC. In addition Caspase 8 and Bim can be activated by TNF through a pathway of activation and inhibition events. A key regulator of both Caspase 8 and Bim activation is the NFκB module, which is being activated by TNF. This module results in mRNA transcription and translation of a protein P which functions as a buffer for Reactive Oxygen Species (ROS), modelled to appear after a certain time as a result of TNF activation. If ROS is not sufficiently buffered it will result in increased translation of JNK, which is activated by upstream signals of the TNF activation pathway. The activated JNK acts by both directly activating Bim and by inhibiting inhibitors of Caspase 8 activation.

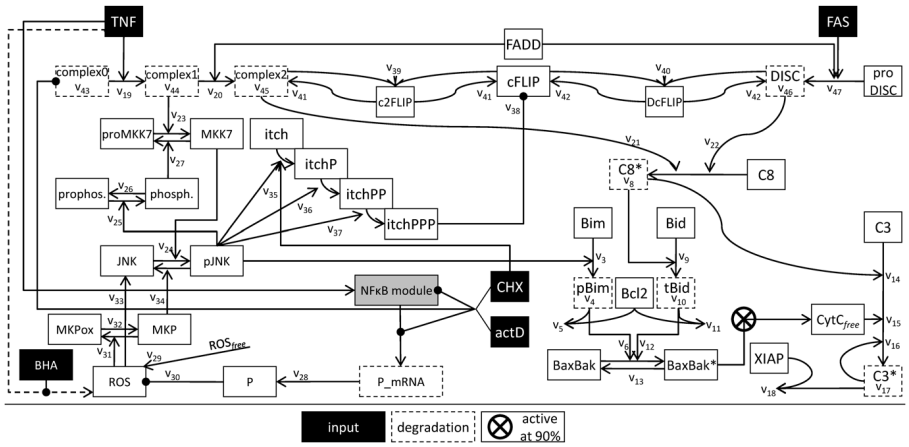


Figure 3.8: Illustration of the interactions in the larger apoptosis model (Illustration was published by Schlatter et al. [91] under Creative Commons Attribution (CC-BY) license. doi: 10.1371/journal.pone.0018646.g005).

All these interactions yield a very complex network, but the end result is similar to that in the smaller apoptosis mode, a burst of activated Caspase 3, leading to apoptosis.

A key difference from the smaller model, is that many proteins in the larger models do not have any production and degradation modelled. Instead the concentrations are set as initial variables and stay constant throughout the model, or decrease through degradation. This means that a parameter sensitivity analysis will not be able to capture the effect of SNPs affecting expression levels of these proteins.

3.1.5 Work covered in this chapter

In this chapter two main questions will be asked:

1. Which parts of the apoptosis pathway are more sensitive to perturbations than others?

2. How do these sensitivities change when an individual acquires mutations over the course of their life?

To answer these questions the two apoptosis models mentioned earlier will be studied using two sensitivity analysis tools as well as parameter/variable scans with in-house scripts.

The idea will be that the perturbations used in the sensitivity analysis tools will correspond to a random set of SNPs affecting either the concentration or function of a protein in the model. By studying the sensitivity patterns from the two methods, the first question can be studied.

By introducing initial perturbations of the system, corresponding to somatic mutations, and determining the sensitivities after these perturbations, the second question can be studied.

On their own, the two sensitivity analysis tools will only reveal how sensitive the dynamics of the system are to perturbations. This will then have to be related to the biological question of when a SNP would push the cell over from a normal cell, committing to apoptosis upon activation, to one which does not commit to apoptosis. To do this the two phenotypes will have to be interpreted in terms of system dynamics. For the two models chosen, the commitment to apoptosis will be interpreted as the activation of Caspase 3 reaching a pre-defined threshold within a given time-frame.

In order to know when the sensitivities of the model can be interpreted as actual potential to push the system over from a normal to an abnormal phenotype, domains of the configuration space will have to be identified, where this is possible. Using parameter or variable scans of the two models these domains will be identified.

3.2 Materials and Methods

3.2.1 Models

Both models were implemented in python 2.7 using the odeint solver from SciPy v.0.16.0. In order to be able to run the sensitivity analysis tool SASSy [101], the models were also implemented in Matlab 8 using the specified format required by the tool. Furthermore, for analysis using the python package SloppyCell [102], the models were also implemented using the format required for that tool.

3.2.1.1 Smaller Apoptosis Model

A small apoptosis model published by Eissing *et al.* (2004) [95] was implemented as described in the paper (Figure 3.7). The model takes a Caspase 8 activation signal as input, which has been set to 1,000 molecules if not stated otherwise. All equations, parameter settings and initial conditions can be found in Equation B.1-B.8 and Table B.1 and B.2.

In order to be able to easily compare the time to apoptosis between different runs of the model, the time at which the active form of Caspase 3 reaches its maximum has been chosen to be interpreted as the time of onset of apoptosis. This cascade of Caspase 3 activation is the last event which occurs in the model upon upstream activation, although, biologically, there are further downstream reactions involved in apoptosis. Biologically, a signal will need to have a certain strength to give any significant downstream results. In order to take this into account, a condition was applied, that the activation would have to reach at least 1000 molecules of Caspase 3 in order to be considered a commitment to apoptosis. The model was run up to 5000 minutes, if not stated otherwise.

3.2.1.2 Larger Apoptosis Model

A model of TNFα induced apoptosis signalling published by Schlatter *et al.* [91] was implemented as described in the paper (Figure 3.8). The node corresponding to translational inhibition by cycloheximide was always set to 0, as was the node corresponding to the antioxidant butylhydroxyanisol (BHA), as well as the node corresponding to translational inhibition by actinomycin D. The activation signal coming from TNFα was set to 100 and FasL was set to change from 0 to 100 after 12 hours, as described in the paper. All equations and parameter settings can be found in equation C.1-C.47 and Table C.1 and C.2. The onset of free ROS production was modelled by the function:

$$ROS_{free}(t) = \frac{1}{0.03 \times 2\pi} e^{\frac{1}{2}(\frac{t-4}{0.03})^2} \times 100 \times (1 - BHA) \tag{3.9}$$

resulting in a short burst of 100 ROS units being released after 4 hours. The model was run for 20 hours unless stated otherwise.

3.2.2 Variable and parameter scan

3.2.2.1 Smaller Apoptosis Model

To assess how sensitive the model is to changes in initial concentrations of its components the production parameters were changed between 0% and 200% of published values. When parameters were perturbed from the standard values, the system was first run for 5000 minutes with initial Caspase 8 activation set to zero to allow the system to find its new steady state. The final concentrations in this steady state were then used as initial concentrations in the actual run, which included different amounts of initial activation of Caspase 8. The model was then run for 20,000 minutes in steps of 1.0 minute.

For each set of runs the maximum concentration of active Caspase 3 and time to apoptosis was recorded.

3.2.2.2 Larger Apoptosis Model

To assess the sensitivity in a comparable way to that of the smaller model, each variable that was not 0 at time = 0, was separately changed incrementally from 0% to 200% of published values [91] in steps of 10% while the other initial concentrations were maintained at their original values.

The models were then run from time 0 to 20 hours and the output reported in steps of 0.2 hours in order to get a high resolution dataset of the dynamics. Within that time frame the maximum activation signal was measured as well as time to activation, calculated as time to maximum peak of Caspase 3 activation.

3.2.3 SASSy

To assess the sensitivity of the system with regards to the entire parameter set the program SASSy was used [101].

3.2.3.1 Smaller Apoptosis Model

As an input for SASSy, the model was implemented in Matlab 8 with the same initial values and parameter settings as stated in Table B.1 and B.2. It was run with either

500 or 3,000 molecules of active Caspase 8 at starting time and analysed from time 0 to 4,000 minutes. In addition the model was analysed from time 0 to time of the maximum peak of activated Caspase 3 (manually estimated to 2,003.2 and 328.6 for 500 and 3,000 molecules, respectively) and within a short window around the maximum of activated Caspase 3 (manually set to 1,800.1–2,100.2 minutes and 36.2–500.8 minutes for 500 and 3,000 molecules, respectively).

Time course dynamics, Singular values and PC vectors were recorded.

3.2.3.2 Larger Apoptosis Model

The model was implemented in Matlab 8 with the same initial values and parameter settings as stated in Table C.1 and C.2. The binary parameters corresponding to presence of cycloheximide, Actinomycin D, BHA and TNFα and Fas signalling were set to be invariable as to not be included in the sensitivity analysis. The model was run to 20 hours and time course dynamics, Singular values and PC vectors were recorded.

3.2.4 SloppyCell

SloppyCell is a sensitivity method which calculates average square change in node values over time in order to quantify the change in model dynamics as parameters are perturbed from their initial values [102]. The details of how it works are outlined in Section 3.1.3.1.2.

All sensitivity analysis using SloppyCell in this thesis was conducted as outlined in Algorithm 1.

Algorithm 1 SloppyCell Sensitivity Analysis

1: Load SloppyCell Reaction Network RN of the dynamical system
2: Alter any initial conditions if necessary
3: Calculate time window by integrating system and measure time of maximum activated Caspase 3
4: Set time window for analysis
5: Set variables to track as "experimental" data. (Used to calculate Hessian of perfect data)
6: Calculate Sensitivity trajectory from RN
7: Calculate Hessian of perfect data w.r.t. $\log \theta$ (Equation 3.8)
8: Perform PCA on Hessian

3.2.4.1 Smaller Apoptosis Model

The model was analysed as described in Algorithm 1 with the following configurations of initial conditions and experimental data:

1. Values from all eight nodes in the model were used as experimental values and all parameters were set as optimisable.
2. Just the values of active Caspase 3 were used as experimental values and all parameters were set as optimisable.
3. Values of active Caspase 3 were used as experimental values and the production rate parameters were set as optimisable.

The analysis was performed from time zero and stopped when the activation peak of Caspase 3 reached its maximum, for each instance of the simulation individually, or when the simulation reached the end time. The level of initial Caspase 8 activation varied between 500, 1,000 and 3,000 molecules, representing changes in upstream pathways. Eigenvalues and eigenvectors for all parameters, as well as time course data for all components in the model were plotted. Additionally, the Hessians with regards to one single parameter indicating single parameter sensitivity were also plotted.

3.2.4.2 Larger Apoptosis Model

The method was applied to the larger apoptosis model. All parameters were set as optimisable except: actD, TNF, BHA, CHX, Tr, and CytC_{free}, which were set to be fixed. The model was then analysed using just Caspase 3 as experimental values. The model was run to time 20 hours. Eigenvalues and eigenvectors for all parameters, as well as time course data for all components in the model were plotted. Additionally, the Hessians with regards to one single parameter indicating single parameter sensitivity were also plotted.

3.3 Results

3.3.1 Smaller Apoptosis Model

3.3.1.1 Model Behaviour

When running the model under normal conditions, with an initial activation of 3,000 molecules, the time to commitment to apoptosis, that is the time until a burst of Caspase 3 activation of over 1,000 molecules, was around 350 minutes (Figure 3.9). As expected by the published results [95], an increase in the initial activation signal of active Caspase 8 did not significantly alter the response time of the system, whereas decreases gradually increased the response time. For smaller changes in the activation signal the change in response time was modest. However, with an ever lower activation signal, the response time of the system quickly increased.

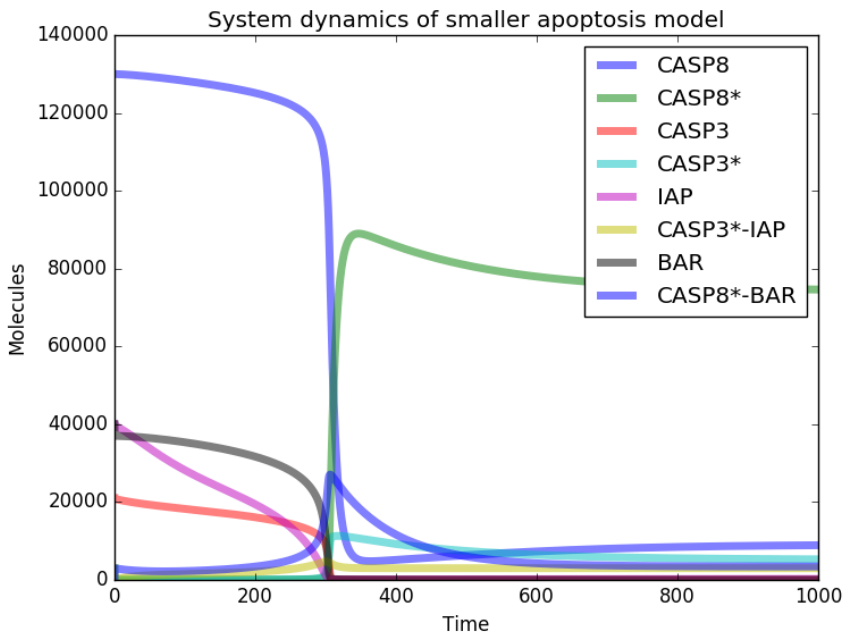


Figure 3.9: Time course dynamics of small apoptosis model. Initially, the concentrations of the active forms of Caspases are held very low due to rapid uptake by the inhibitors. Eventually the activation of the Caspases becomes faster than the renewal of the inhibitors, resulting in an accelerated depletion of the inhibitors and a burst of activation of both Caspases.

3.3.1.2 Parameter Scan shows two distinct types of model behaviour

When keeping the activation signal constant at 3,000 and perturbing two parameters at a time to various extent, the time to reach a peak in Caspase 3 activation (here measured

as time to maximum concentration of active Caspase 3) showed relatively little variation, until it, within a very narrow window, flipped from activating apoptosis within 5,000 time units to not activating apoptosis (Figure 3.10). Even when extending the simulation to 20,000 time units, the same pattern was observed, indicating that the system had, indeed become unresponsive, at least from a perspective of biological relevance (Appendix Figure D.1). Note, that it is not necessarily the case that Caspase 3 activation did occur but did not reach the threshold set before. It is possible that a perturbation of some parameters would cause a smaller peak but not affect the time of the peak very much. However, for example in the case of the production rate of Caspase 8, the system changed behaviour, at least within the time frame set. In the non-responsive configuration there was no activation past a couple of molecules initially (this as also seen in the case of responsiveness) and no extended activation thereafter, whereas in the responsive configuration, the activation was in the range of several thousands of molecules and there was an sustained, lower activation after the initial peak (Figure 3.11, visualised at 500 molecules of initial activated Caspase 8 due to the slower progression at that level). It is possible that the system, even when deemed unresponsive here, would eventually yield a peak in activated Caspase 3. Indeed, by again perturbing the production rate of Caspase 8 it was seen how the the activation signal was sustained, and gradually moved up to over 25,000 time units (Figure 3.12, 500 molecules of initial activated Caspase 8). However, after such a long time, the cell would have undergone cell division already and even if the mathematics indicate that the system would activate apoptosis, it would not be relevant from a biological point of view. Likewise, even if the threshold for activation is not known and had to be assumed in this work, the concept of a signal needing to reach a certain level of sustained activity before being biologically relevant is consistent with how many of biological signalling systems work.

By changing the amount of initial activation signal, the location, and to some extent the shape, of the border, at which activation no longer occurred above the given level and within the set time frame, shifted. The sensitivity was lower for lower amounts of activation signal. For example, when using 500 instead of 3,000 molecules as starting activation, the system was already in the area where changes in activation signal had a significant effect on the response time of the system. This also meant that the amount of perturbation in any parameter required to flip the system to not respond within 20,000 minutes (to not yield a signal above the threshold with a subsequent sustained signal) was much smaller than when using 3,000 molecules as activation signal. However, once the system was in the area of sensitivity, a much smaller additional perturbation was required

to flip the system if 3,000 units of activated Caspase 8 was used, compared to 500 units (Figure 3.13) This was true for all parameters under investigation (Appendix Figure D.2).

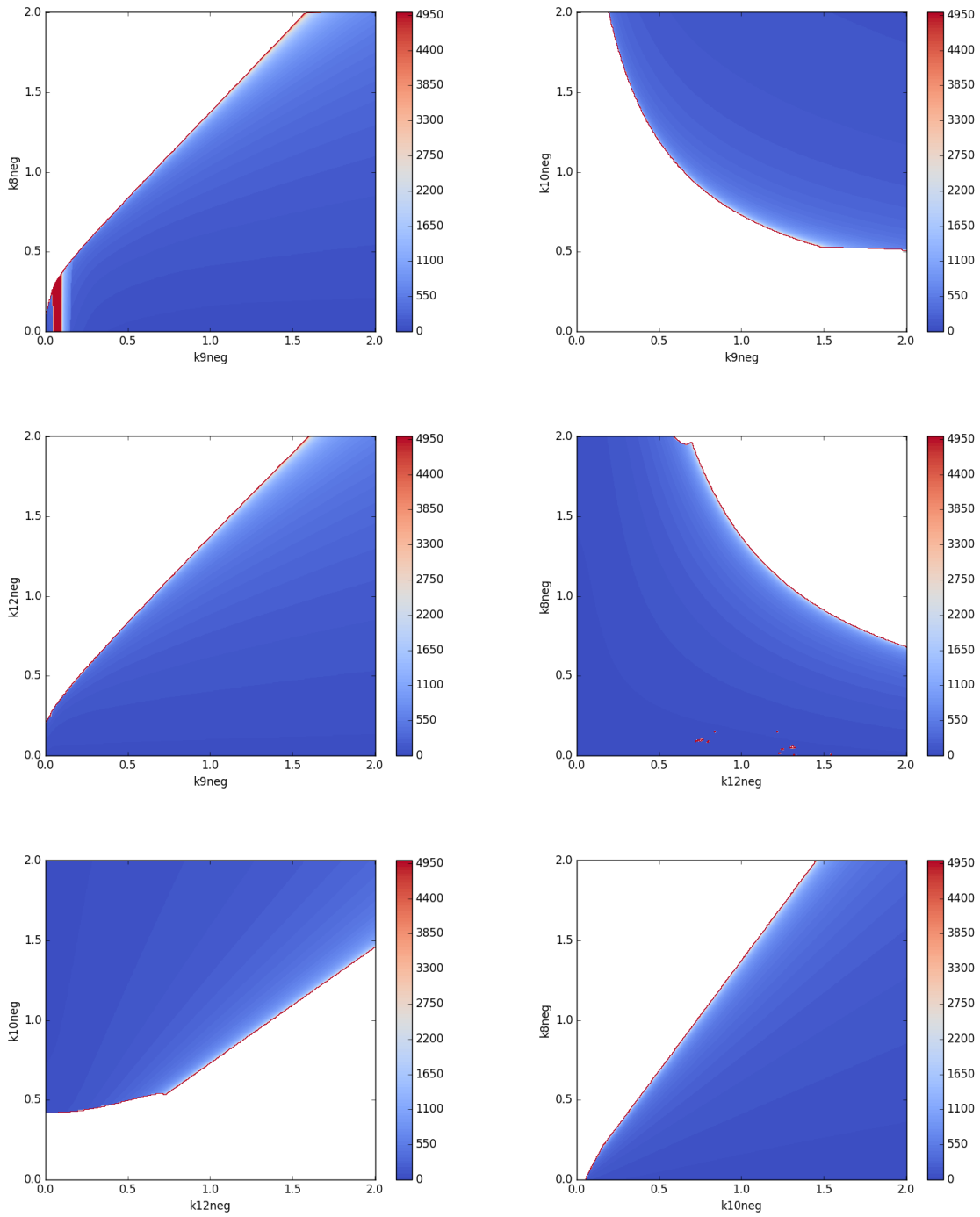


Figure 3.10: Time to maximum Caspase 3 signalling when perturbing two parameters between 0 and 2 time the initial value. The value of the parameters is depicted on the respective axis and time is colour coded from 0 (dark blue) to 5,000 (dark red). Upon small perturbations the time to apoptosis does not alter much. However, within a very small window of parameter perturbation the time changes from very short to very long. The white indicates where time to apoptosis exceeded the limit of the scale.

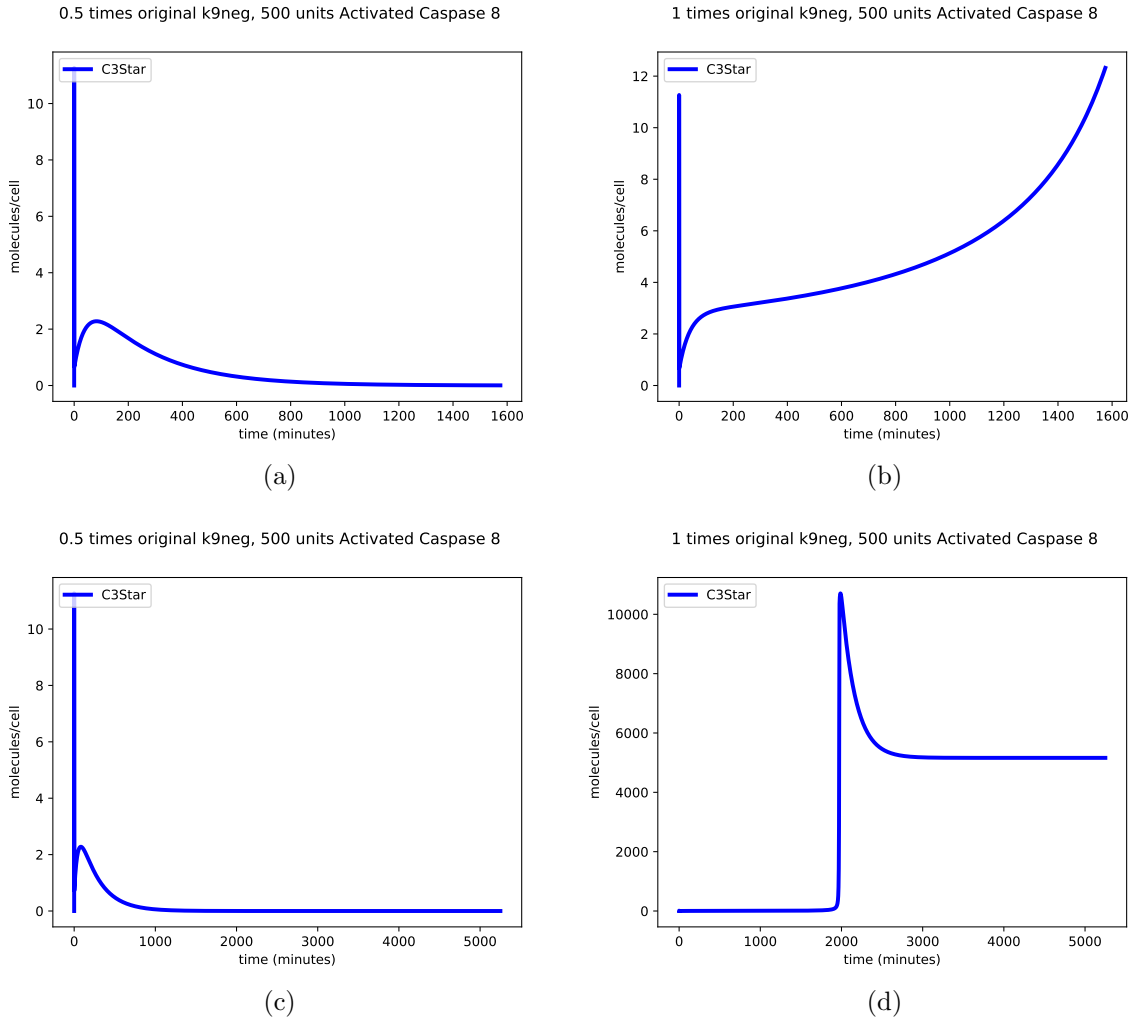


Figure 3.11: Difference in activated Caspase 3 trajectory for responsive and non-responsive systems. When the system is deemed non-responsive there is a very small initial peak in activated Caspase 3, which quickly dies out (a and c). When the system is deemed responsive (b and d) there is again a small initial peak. However, the activation then builds up until there is a large activation peak followed by a sustained activation.

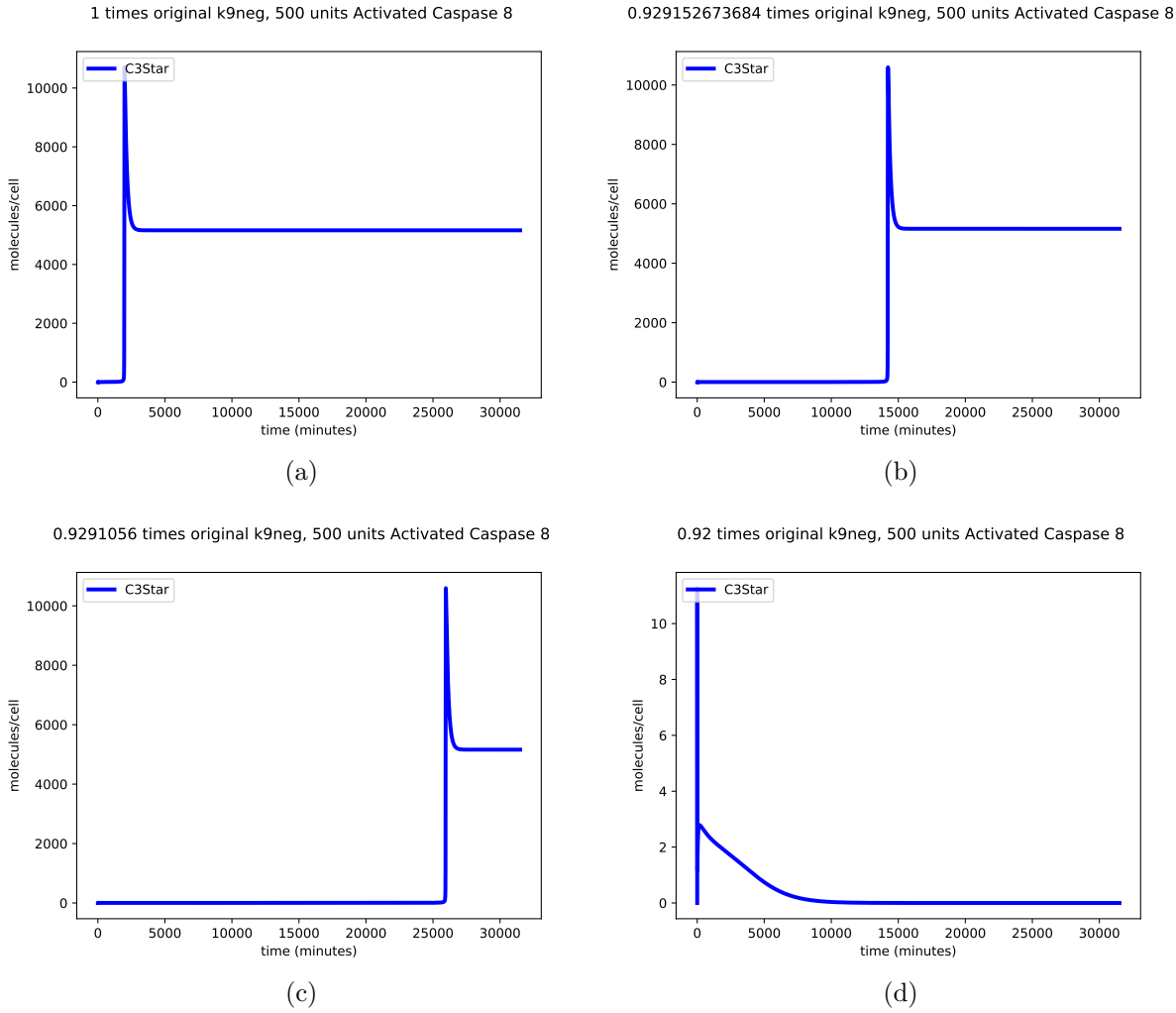
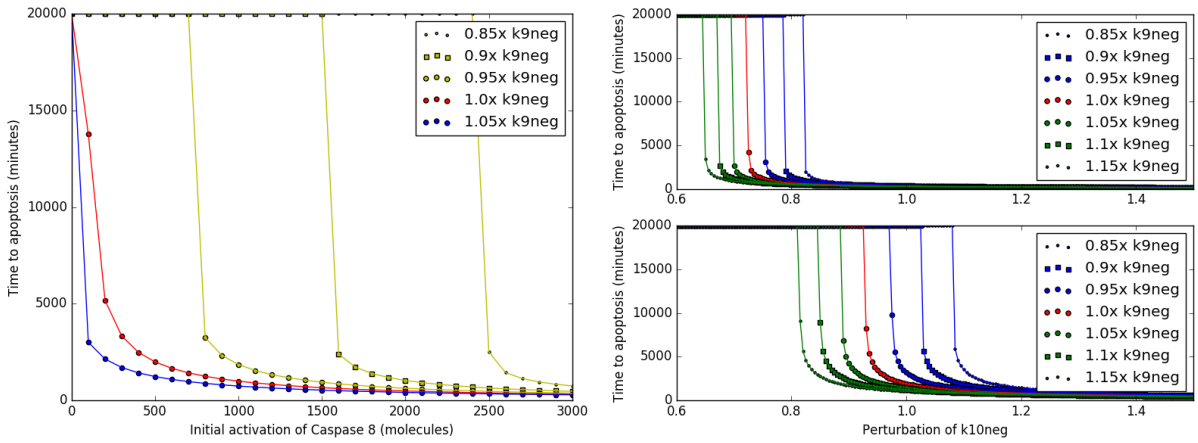


Figure 3.12: As the production rate parameter of Caspase 8 is perturbed from initial value (a) gradually down to 0.92 (b, c, and d) the time of the large activation peak shifts upwards until it eventually disappears outside of the time frame of the simulation. The simulation was performed with 500 molecules of initial activated Caspase 8 due to the slower shift in time to onset of apoptosis after perturbation of the parameter.



(a) (b)

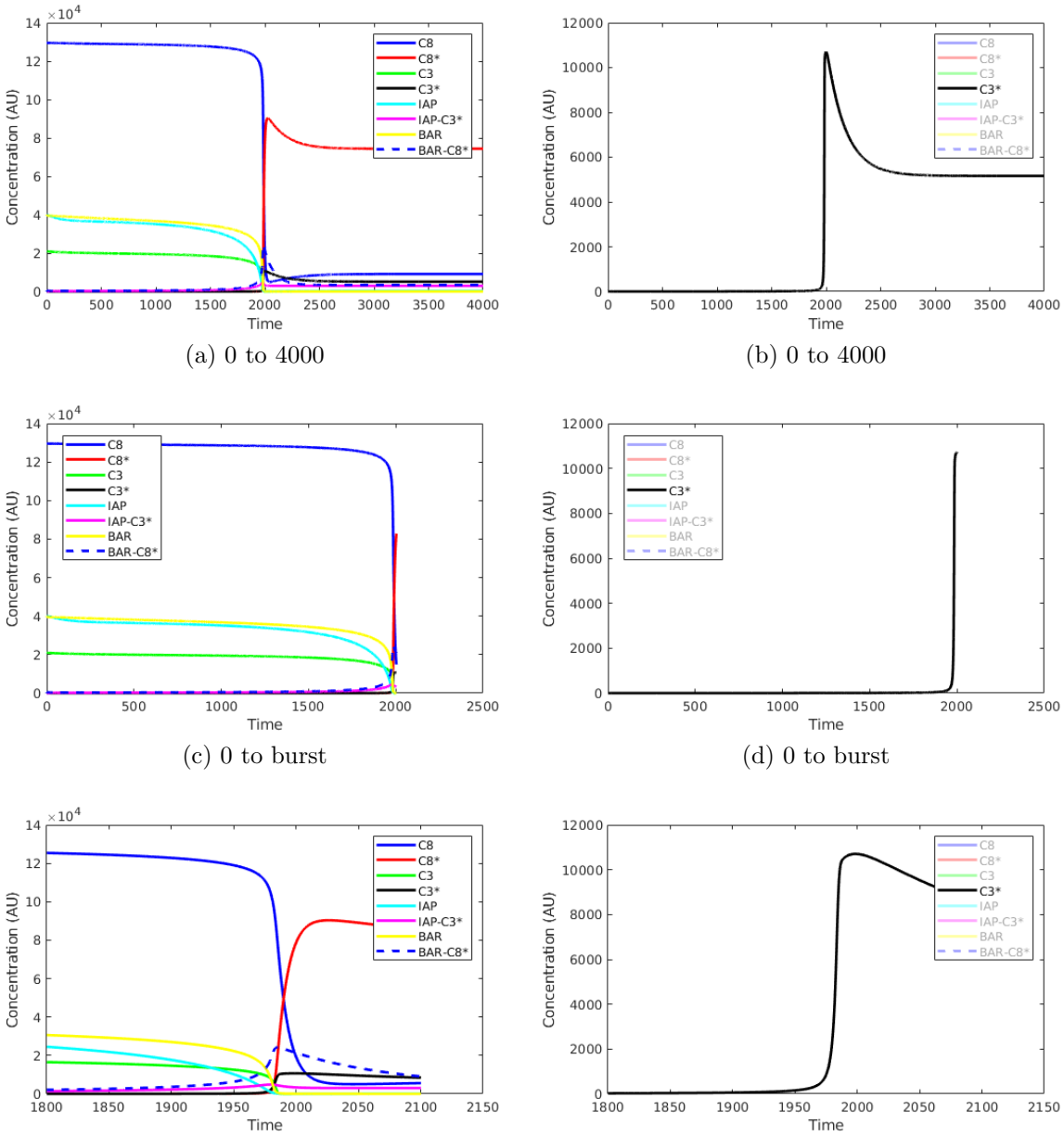
Figure 3.13: a) Time to apoptosis as a function of initial active Caspase 8 signalling. Red lines show standard parameter setting. Blue lines show the same pattern with an increase of $k9_{-1}$ (corresponding to production rate of Caspase 8) of 5%. Yellow lines show the trend with a decrease of 5, 10 and 15%. Initial settings indicate a window of Caspase 8 activation signalling, in which the time to apoptosis quickly increases and eventually exits the time frame of the simulation. The blue and yellow lines indicate a shift of this signalling window. b) Time to apoptosis as a function of parameter $k10_{-1}$. Blue and green lines show an additional perturbation of parameter $k9_{-1}$ (in increments of 5%), whereas red shows the initial value. Top and bottom plots show pattern when initial Caspase 8 activation is set to 3000 and 500, respectively. The system shows a similar window a parameter perturbations as was seen when altering Caspase 8 activation, where the system response time rapidly increases. The location of this pattern proves to depend both on additional parameter perturbations and the amount of initial Caspase 8 activation.

3.3.1.3 SASSY indicates that most of the sensitivity is centred around the time of behaviour switching

Using the program SASSy the small apoptosis model was analysed using two different amounts of initial activation signal: 500 and 3,000, and three different time settings. First the model was analysed from time 0 to 4,000 minutes, then it was analysed to the time of maximum concentration of activated Caspase 3. Lastly it was analysed within a window of 400 minutes around the maximum concentration of activated Caspase 3. The dynamics of all settings are depicted in Figure 3.14 and 3.15 for 500 and 3,000 molecules of activated Caspase 8, respectively.

Parameter sensitivity spectra were generated for all cases analysed. Normalised singular spectra were summarised in Figure 3.16 and 3.17 for 500 and 3,000 molecules of activated Caspase 8, respectively. There was very little variation in the spectra between the three time frames in both sets of Caspase 8 activation and in all cases there was a rapid decline in importance among the Singular values. When using 500 molecules there were only 2 singular values within 1% of the largest value. Using 3,000 molecules, there was slightly slower decay, with 4 values within 1% of the largest value.

When looking at the parameter spectra for each principal component the pattern between the three windows of analysis was also very similar across all PCs (Figure 3.18). This was true for both amounts of Caspase 8 activation. In all cases the higher PCs were dominated by two parameters, which became even more clear, when looking at them separately (Figure 3.19). The four most important parameters turned out to be the production parameters for the four proteins in the model.



(a) 0 to 4000

(b) 0 to 4000

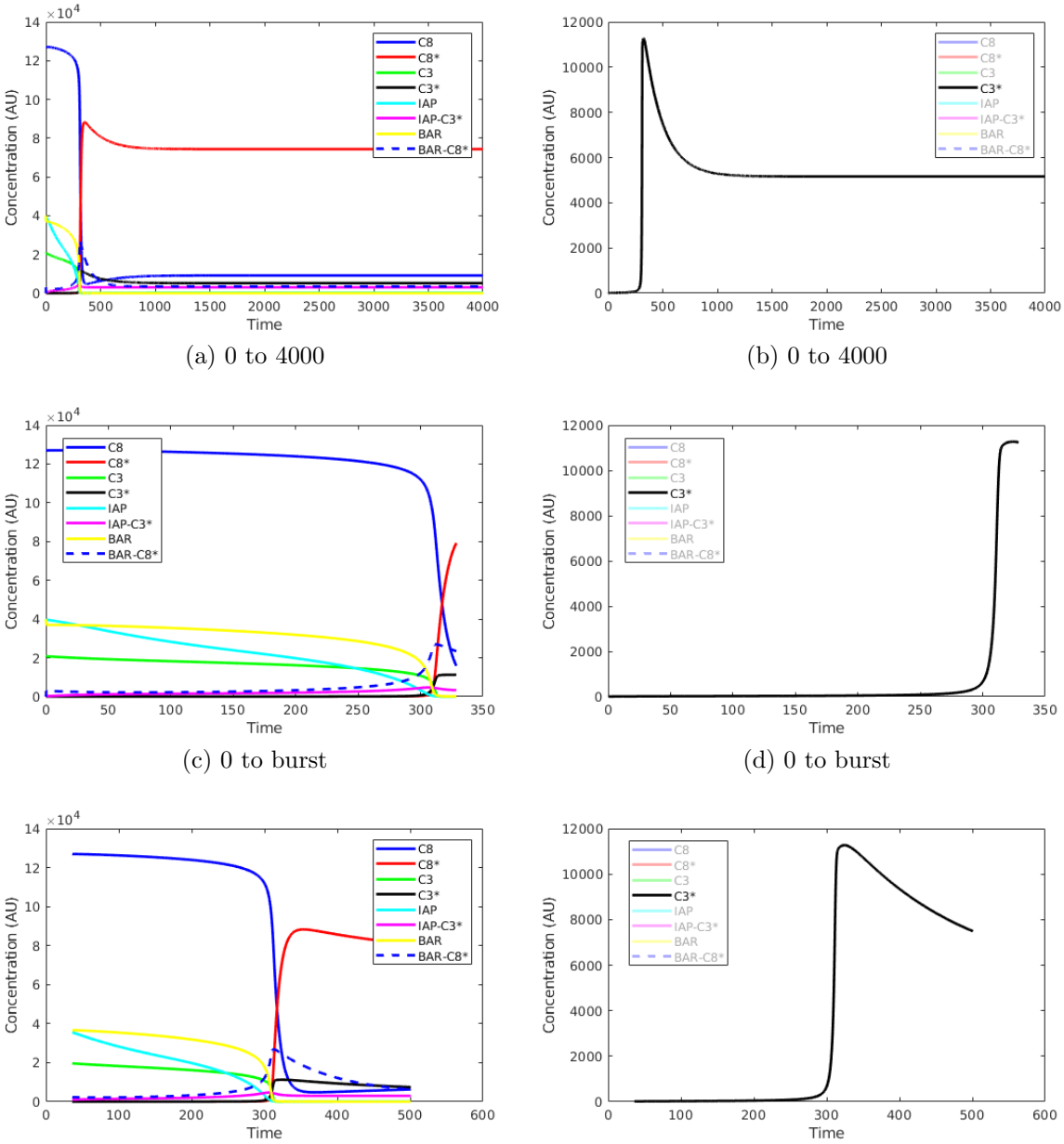
(c) 0 to burst

(d) 0 to burst

(e) burst

(f) burst

Figure 3.14: Time course data for all the three sets of analyses performed on the smaller apoptosis model using 500 as initial Caspase 8 activation signal (as opposed to 3000 in Figure 3.15). In all three cases the same parameter settings were used, but different time frames were analysed. (a) and (b) show dynamics from time 0 to 4000, (c) and (d) from 0 to peak of Caspase 3 activation and (e) and (f) within a window around the Caspase 3 activation burst. (a), (c) and (e) show the dynamics for all 8 variables, whereas (b), (d) and (f) show the dynamics of only active Caspase 3.



(a) 0 to 4000

(b) 0 to 4000

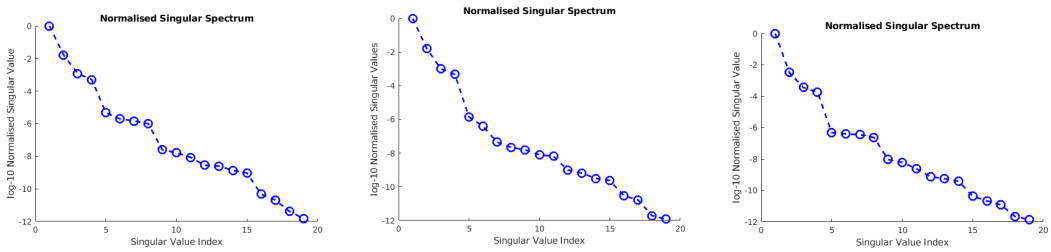
(c) 0 to burst

(d) 0 to burst

(e) burst

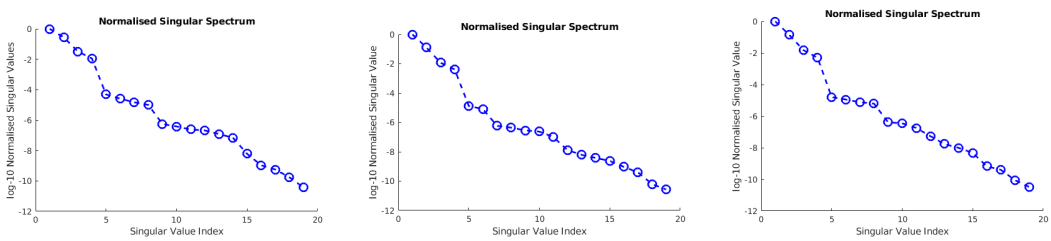
(f) burst

Figure 3.15: Time course data for all the three sets of analysis performed on the smaller apoptosis model using 3000 as initial Caspase 8 activation signal (as opposed to 500 in Figure 3.14). In all three cases the same parameter settings were used, but different time frames were analysed. (a) and (b) show dynamics from time 0 to 4000, (c) and (d) from 0 to peak of Caspase 3 activation and (e) and (f) within a window around the Caspase 3 activation burst. (a), (c) and (e) show the dynamics for all 8 variables, whereas (b), (d) and (f) show the dynamics of only active Caspase 3.



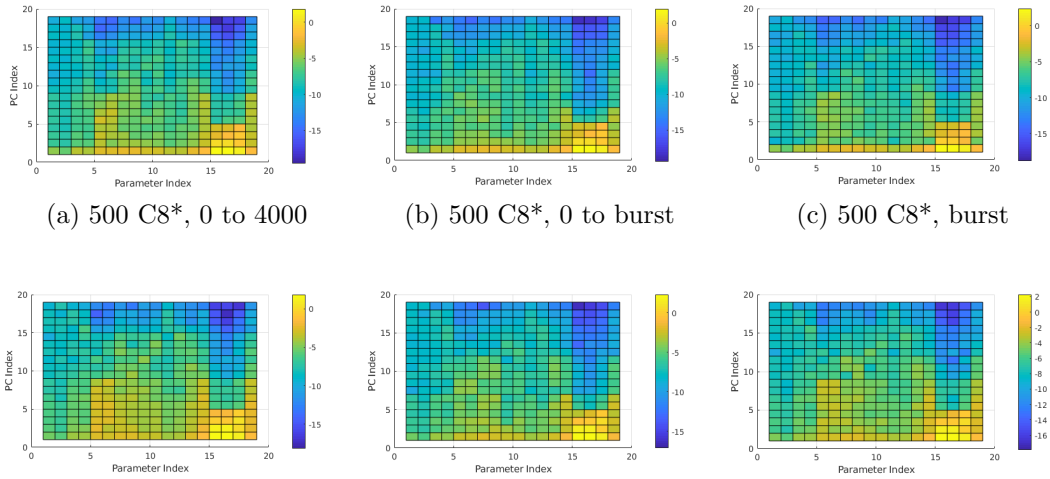
(a) 0 to 4,000 (b) 0 to burst (c) burst

Figure 3.16: Singular spectrum for SASSy analysis of the smaller apoptosis model, using 500 AU as initial active Caspase 8 concentration and the time frames: (a); from 0 to 4000, (b); from 0 to peak of active Caspase 3 concentration and (c); a window of about 400 time units around the maximum activated Caspase 3 concentration. All three analyses show a similar pattern of rapidly declining Singular values and the first Singular values being 2 orders of magnitude higher than the second.



(a) 0 to 4,000 (b) 0 to burst (c) burst

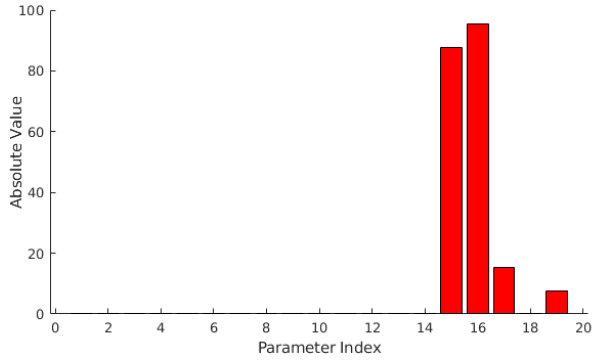
Figure 3.17: Singular spectrum for SASSy analysis of the smaller apoptosis model, using 3000 AU as initial active Caspase 8 concentration and the time frames: (a); from 0 to 4000, (b); from 0 to peak of active Caspase 3 concentration and (c); a window of about 400 time units around the maximum activated Caspase 3 concentration. All three analyses show a similar pattern of declining Singular value. However, the decline is not as fast as when using 500 AU of activated Caspase 8 (Figure 3.16) and the first four Singular values are more or less within 2 orders of magnitude.



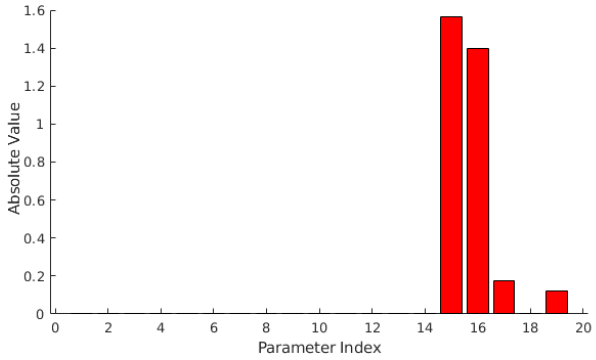
(a) 500 C8*, 0 to 4000 (b) 500 C8*, 0 to burst (c) 500 C8*, burst

(d) 3,000 C8*, 0 to 4000 (e) 3,000 C8*, 0 to burst (f) 3,000 C8*, burst

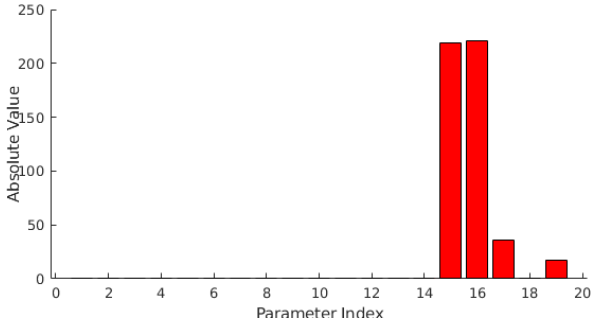
Figure 3.18: Principal component spectrum for SASSy analysis of the smaller apoptosis model using either 500 molecules ((a), (b) and (c)) or 3,000 molecules ((d), (e) and (f)) as initial active Caspase 8 concentration and the time frames: (a) and (d); from 0 to 4,000, (b) and (e); from 0 to peak of active Caspase 3 concentration and (c) and (f); a window of about 400 time units around the maximum activated Caspase 3 concentration. The strength of each parameter has been colour coded according to their $\log_{10}$ absolute values. The parameters occur in the same order as they are listed in table 3.3. All experiments yielded a striking similarity in composition and strength of each parameter, on the x-axis, in the various PCs, on the y-axis. This similarity was maintained both among the three time frames and between the two initial activation signals



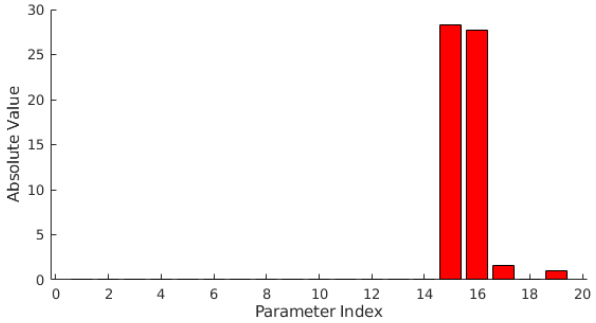
(a) PC1, 500 C8*



(b) PC2, 500 C8*



(c) PC1, 3,000 C8*

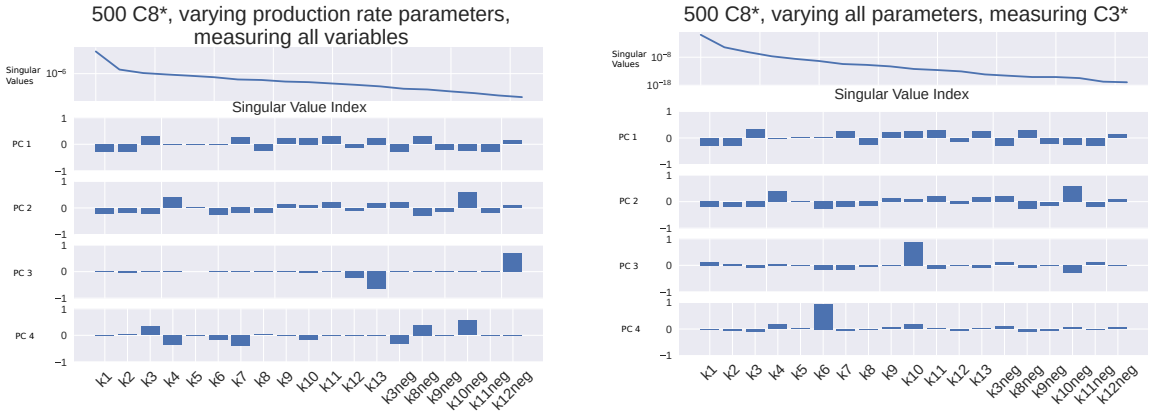


(d) PC2, 3,000 C8*

Figure 3.19: Decomposition of the first ((a) and (c)) and second ((b) and (d) PC using either 500 molecules ((a) and (b)) or 3,000 molecules ((c) and (d)) of initial activated Caspase 8 signal. The parameters occur in the same order as they are listed in Table B.1. Using both amounts of initial activation signal both PCs are dominated by the 15th and 16th parameter, corresponding to production rates of Caspase 8 and IAP, with a smaller contribution of the 17th and 19th parameter, corresponding to production rates of Caspase 3 and BAR. All other parameters had a negligible contribution to the two PCs.

3.3.1.4 SloppyCell analysis reveals slight differences in sensitivity pattern depending on initial conditions

When considering perturbations in all parameters and measuring the sensitivity of the model between time zero and the maximum height of the Caspase 3 peak, there was initially very little difference in the pattern if data from all nodes were used as experimental data, or just that of active Caspase 3 (Figure 3.20). The difference became more clear from the third Principal Component (PC) on. In both cases the first eigenvalue was much higher than the rest and the decay in importance was much slower from the second eigenvalue. Although the first eigenvalue was higher than the rest, there was no clear cut-off between important and unimportant PCs. This was especially true when using only activated Caspase 3 as experimental data, where the difference between the first and the second eigenvalue was much smaller.



(a) All Parameters, All Nodes

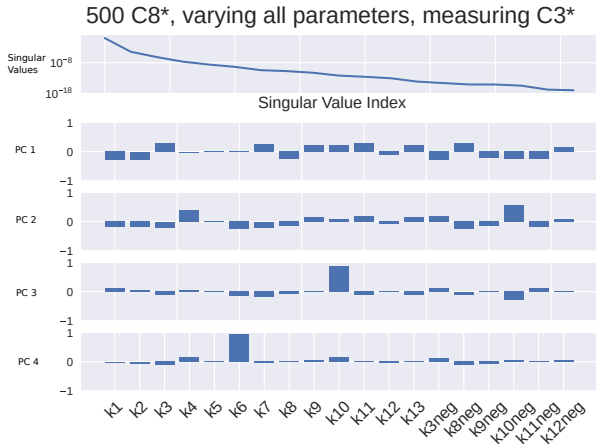
(b) All Parameters, Only Active Caspase 3

Figure 3.20: Left: analysis using all parameters as experimental data. Right: using only active Caspase 3 as experimental data. In both plots, from the top: eigenvalues, 1st PC, 2nd PC 3rd PC and 4th PC. The difference becomes more clear in the lower PCs. The parameters occur in the same order as they are listed in Appendix Table B.1.

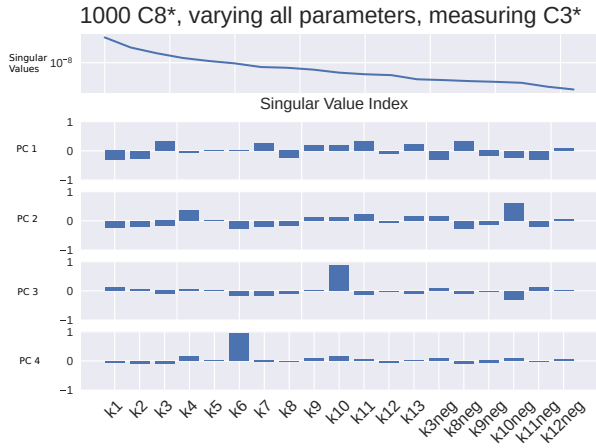
When altering the initial activation signal from 500 molecules of activated Caspase 8 to 1000 and finally to 3000 molecules, there was very little change in the pattern of the PCs (Figure 3.21). However, as the activation signal increased, the dominance of the first PC decreased slightly, indicating a slight difference in sensitivity. In addition to the PCA, the Hessian with regard to single parameter perturbations was also examined. This can be seen as the sensitivity of a single parameter perturbation. When looking at the Hessian for all parameters of the model most parameters had a comparable sensitivity, both when using all variables as experimental data and when using only the amount of

activated Caspase 3 (Figure 3.22). Only a couple of parameters had sensitivity varying by several magnitudes compared to the rest (k4, k5 and k6). As the initial activation signal increased, the system became more sensitive, indicated by smaller Hessians (remember that the hessian approximates an ellipsoid with constant model behaviour change).

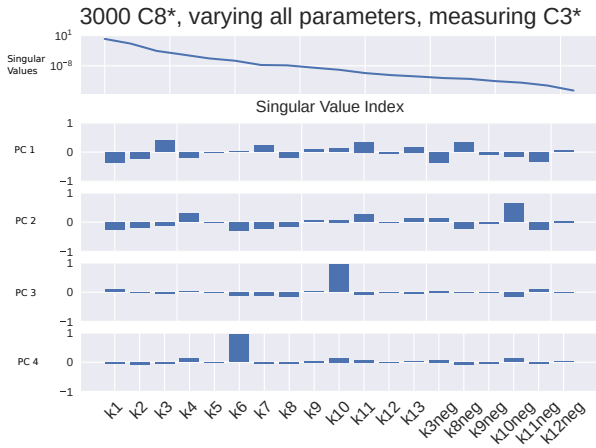
When only taking the production rates into account, again, there was no clear cut-off between important and unimportant PCs (Figure 3.23). For three amounts of activation signal tested the system showed an even spread of sensitivity among all parameters. Again, when looking at the Hessian with regard to single parameter perturbations, all parameters had a similar amount of sensitivity (Figure 3.24). As was the case when analysing the larger set of parameters, the sensitivity increased when the amount of initial activated Caspase 8 increased.



(a) 500 C8*

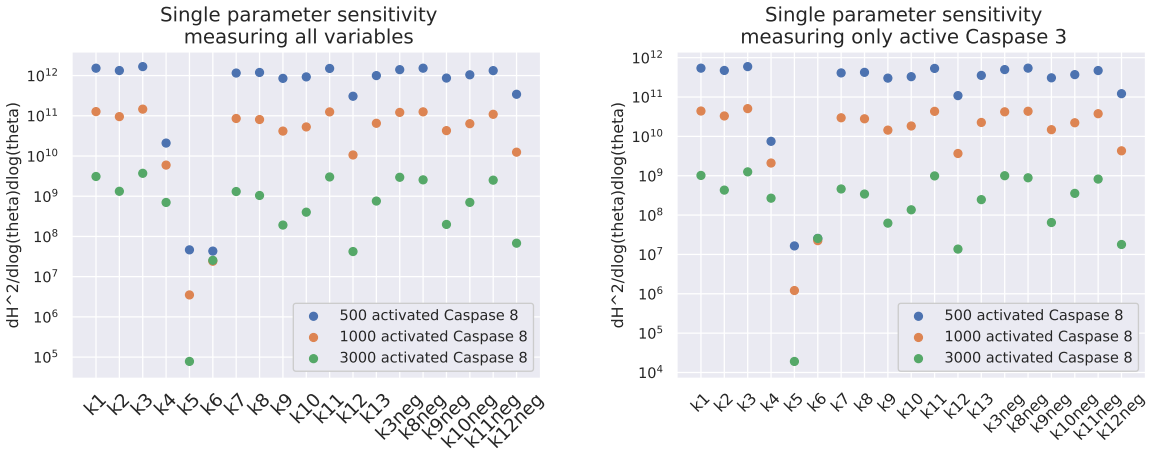


(b) 1000 C8*



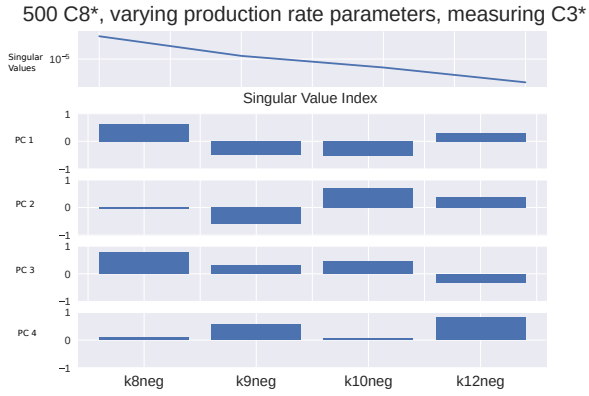
(c) 3000 C8*

Figure 3.21: As the initial activation signal increases from 500 to 1000 to 3000, there are slight differences in the sensitivity patterns of each PC. The parameters occur in the same order as they are listed in Appendix Table B.1.

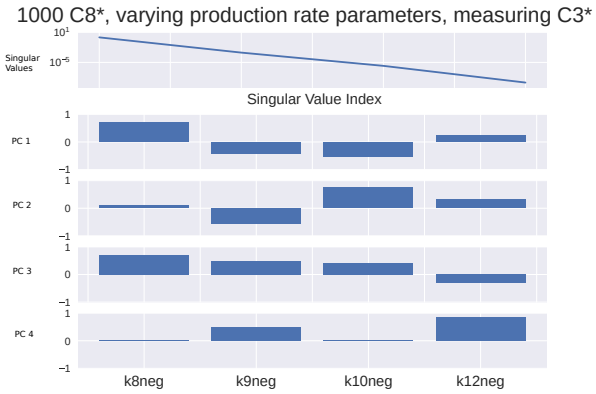


(a) (b)

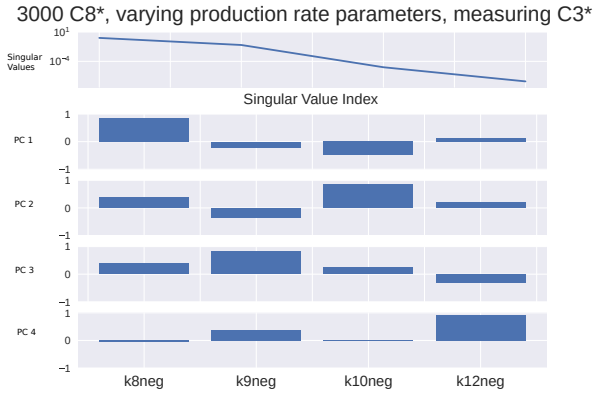
Figure 3.22: SloppyCell Hessian for single parameters ($d\chi^2/d\log\theta_j d\log\theta_j$) in the smaller Apoptosis model looking at all parameters. There is little change between measuring all variables (a) and measuring only active Caspase 3 (b). Increased initial activation signal causes an increase in sensitivity overall (small value indicates high sensitivity as only a small perturbation is needed for the constant model behaviour change)



(a) 500 C8*

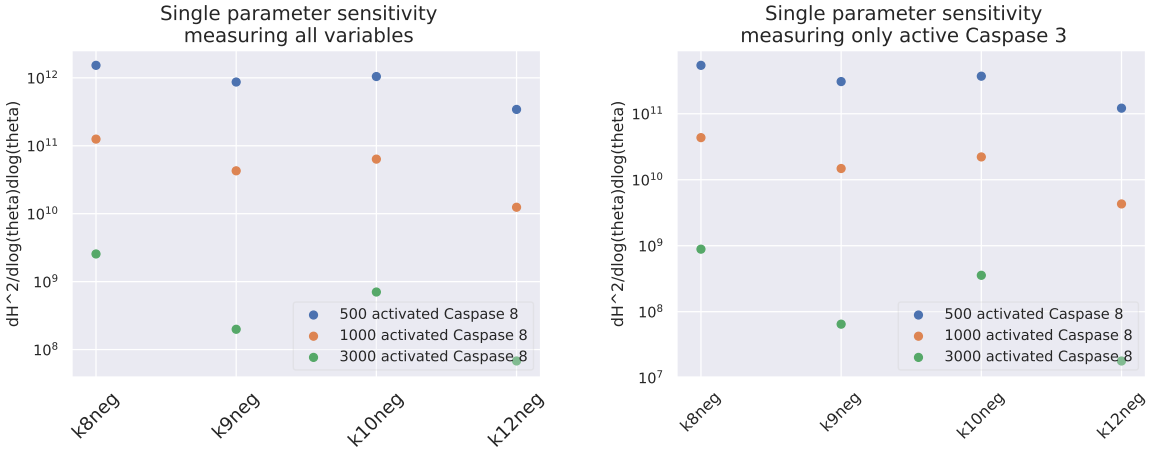


(b) 1000 C8*



(c) 3000 C8*

Figure 3.23: SloppyCell analysis setting only the four production parameters as optimisable and using amount of active Caspase 3 as experimental data. (a),(b) and (c) use 500, 1000 and 3000 as initial Caspase 8 activation respectively. As the initial activation signal increases, there are slight differences in the sensitivity patterns.



(a) (b)

Figure 3.24: SloppyCell Hessian for single parameters ($d\chi^2/d\log\theta_j d\log\theta_j$) in the smaller Apoptosis model looking only at the production parameters. There is little change between measuring all variables (a) and measuring only active Caspase 3 (b). Increased initial activation signal causes an increase in sensitivity overall (small value indicates high sensitivity as only a small perturbation is needed for the constant model behaviour change).

3.3.2 Larger Apoptosis Model

3.3.2.1 Model behaviour

The model was run with a burst of FasL exposure after 12 hours. As expected from the paper, the model did not yield an onset of apoptosis (characterised by a burst of Caspase 3 activation) without FasL exposure or deactivation of protein translation by actinomycin D (not shown). When Actinomycin D was added to the system, Caspase 3 was activated after about 7 hours (not shown), and when FasL was added after 12 hours the system showed a delayed activation of Caspase 3 approximately 13 hours after initiation of the system (Figure 3.25). All results were consistent with previously published results [91].

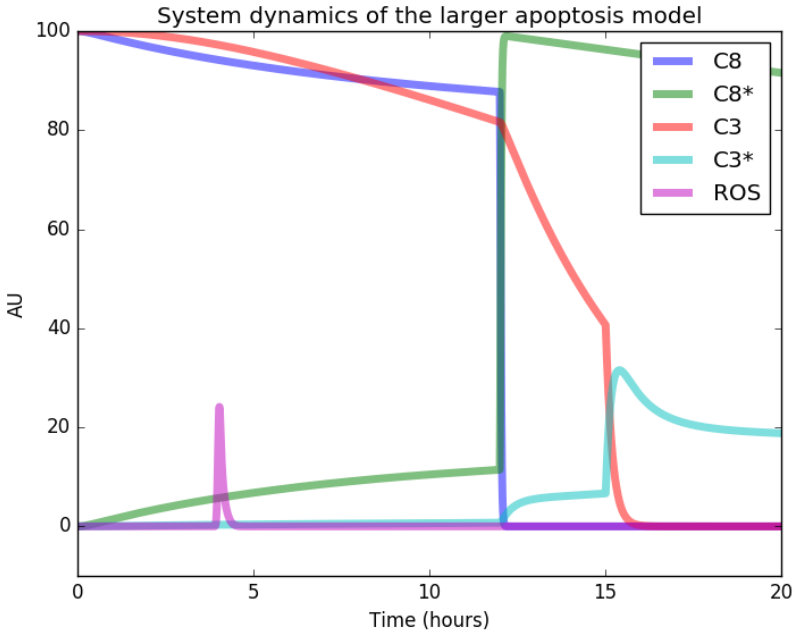


Figure 3.25: System dynamics of the larger apoptosis model. Initially, there is very little Caspase 8 activation and negligible Caspase 3 activation. However, after initiation of Fas signalling there is a spike in Caspase 8 activation at 12 hours, followed by initial smaller activation of Caspase 3. As the inhibitors are exhausted there is a major spike in Caspase 3 activation at around 15 hours, which can be interpreted as commitment to apoptosis.

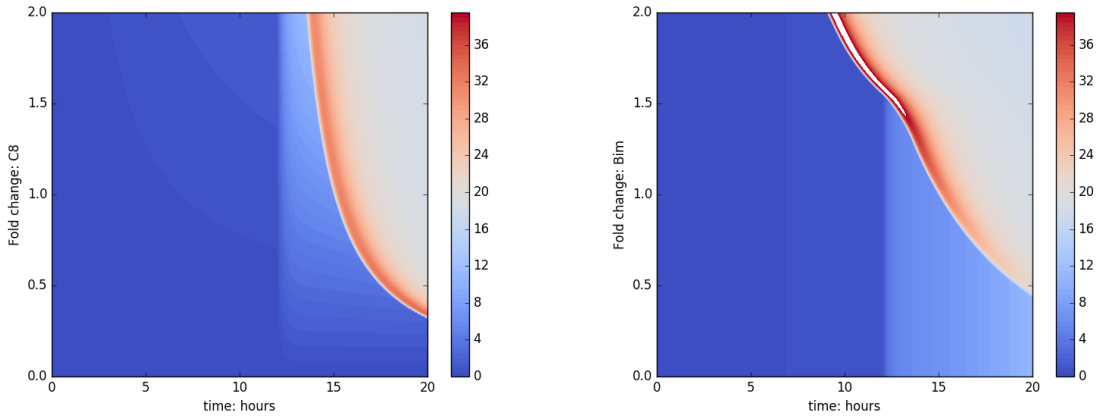
3.3.2.2 Variable Scans reveal sets of characteristic behaviour

Each variable that had an initial value above 0 was varied between 0% and 200% of the initial value, in steps of 1.0%, and the concentration of active Caspase 3 was monitored over time. The results were insensitive to increases in many of the variables and could also withstand changes of up to 50% without losing the apoptotic signalling (Figure 3.26

and Appendix Figure E.1). For many of these proteins, for example Caspase 8 (Figure 3.26a), the time to activation did not vary much when altering the concentration up until a certain point, where the time rapidly increased beyond the time frame of the simulation. Although some proteins had a more gradual change of time to activation, for example Bim (Figure 3.26b), other proteins were very sensitive in a certain direction, for example BaxBak (Figure 3.26c), which could barely take any decrease in concentration before the signal disappeared. Yet other nodes were completely insensitive within the range investigated, for example itch (Figure 3.26d).

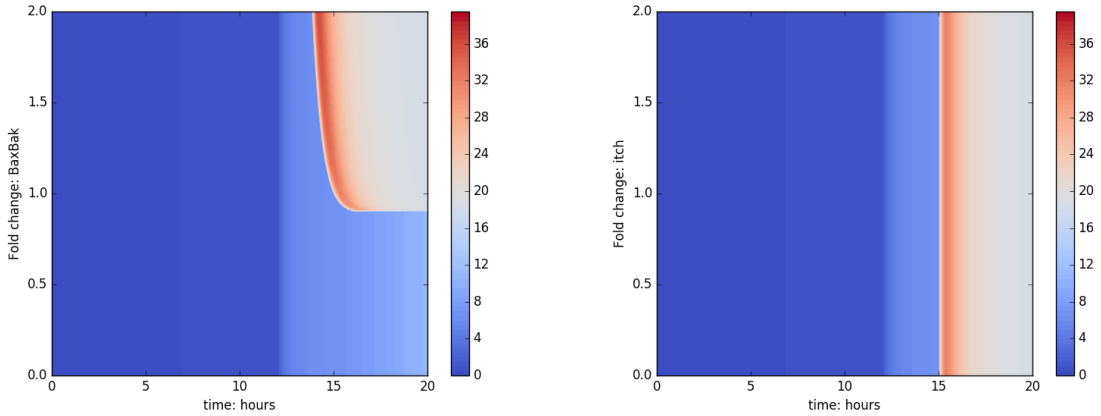
Although most of the proteins maintained an activation signal, which moved in time when altering the concentration, some proteins, for example Caspase 3 and XIAP (Figure 3.27) had a fading strength of maximum activation. This means that it is difficult to establish at what point the system no longer yields a biologically relevant activation signal.

When varying two variables at the same time, the time point at which the system went from high to low final concentration of activated Caspase 3, and the maximum activation of Caspase 3 often depended on the other variable (Figure 3.28 and Appendix Figure E.2-E.13). Some variables, however, did not even have a significant effect on the final concentration of activated Caspase 3 together (Figure 3.29a-d). Still others had value ranges for which the concentration was lower than normal, but not as low as for the other variables, begging the question, as to whether it should be interpreted as an activation signal or not (Figure 3.29e-f).



(a) C8

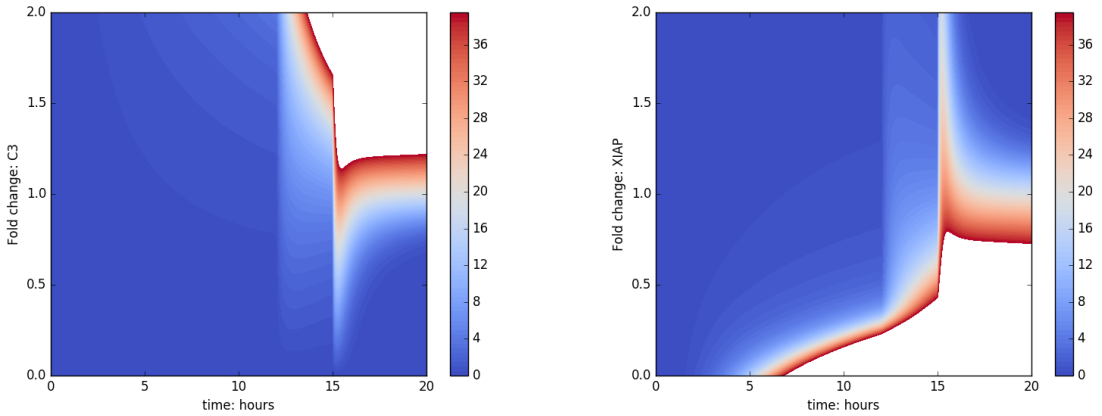
(b) Bim



(c) BaxBak

(d) itch

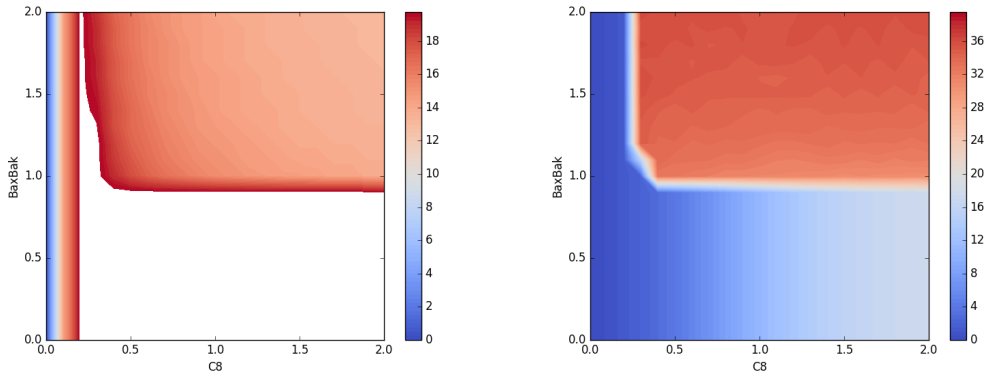
Figure 3.26: Concentration of activated Caspase 3 over time (x-axis) depending on change in initial concentration of one variable (y-axis) between 0 and 2 times the initial concentration. The concentration of Caspase 3 is colour coded between 0 (blue) and 40 (red). Under normal conditions there is a peak in activation around 15 hours after initiation corresponding to the activation of apoptosis. The white indicates where the concentration exceeded the limit of the scale.



(a) C3

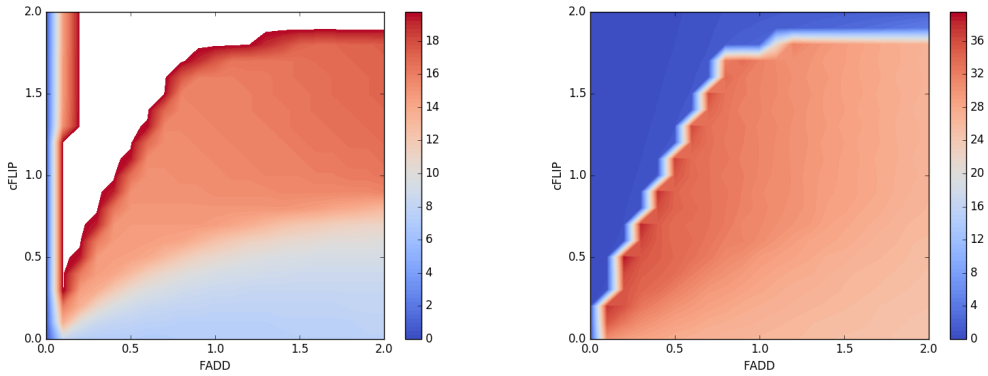
(b) XIAP

Figure 3.27: Concentration of activated Caspase 3 over time (x-axis) depending on change in initial concentration of one variable (y-axis) between 0 and 2 times the initial concentration. The concentration of Caspase 3 is colour coded between 0 (blue) and 40 (red). Under normal conditions there is a peak in activation around 15 hours after initiation corresponding to the activation of apoptosis. The white indicates where the concentration exceeded the limit of the scale.



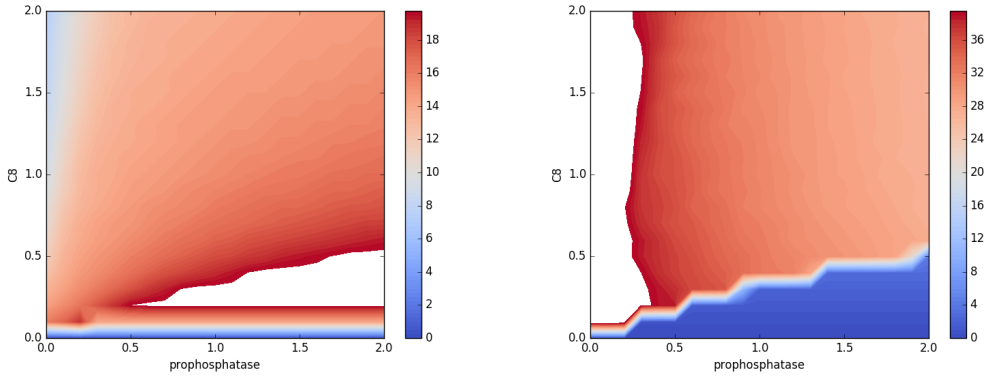
(a) Time to maximum C3*

(b) Maximum C3*



(c) Time to maximum C3*

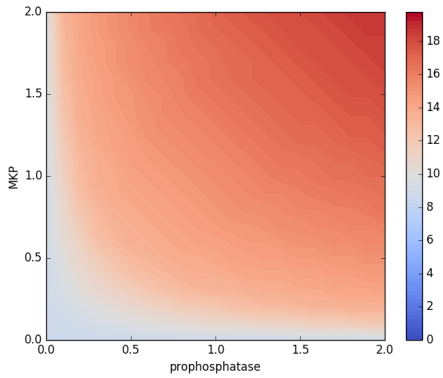
(d) Maximum C3*



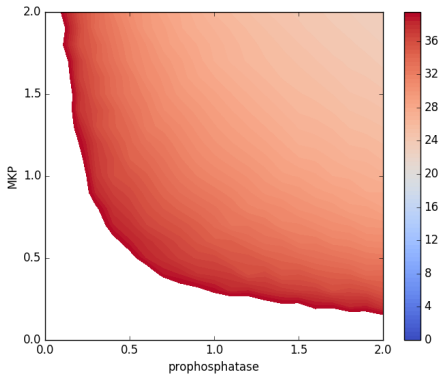
(e) Time to maximum C3*

(f) Maximum C3*

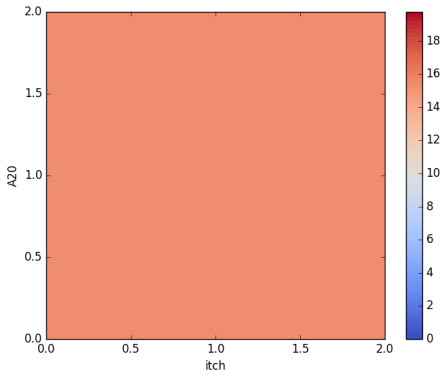
Figure 3.28: Left: Time to maximum Caspase 3 signalling when perturbing two parameters between 0 and 2 times the initial value. The value of the parameters is depicted on the respective axis and the time to maximum Caspase 3 activation is colour coded from 0 (dark blue) to 20 hours (dark red). Right: Maximum Caspase 3 activation upon perturbation of two variables initial values. The maximum Caspase 3 activation ranges between 0 (dark blue) and 40 AU (dark red). The white indicates where the value exceeded the limit of the scale.



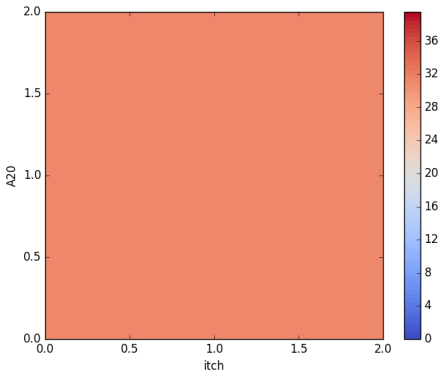
(a) Time to maximum C3*



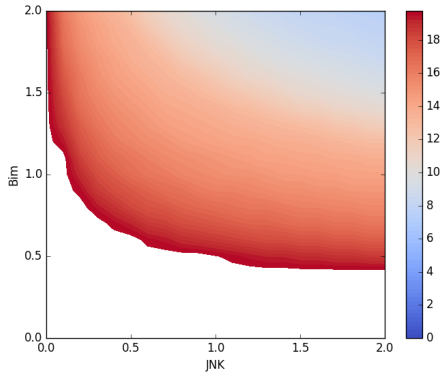
(b) Maximum C3*



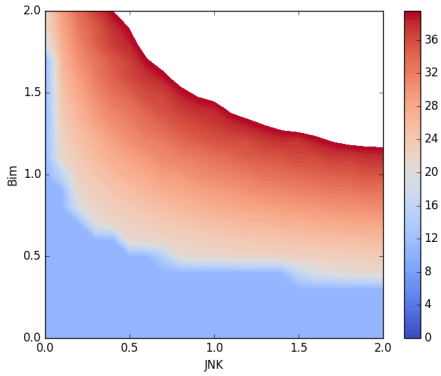
(c) Time to maximum C3*



(d) Maximum C3*



(e) Time to maximum C3*



(f) Maximum C3*

Figure 3.29: Left: Time to maximum Caspase 3 signalling when perturbing two parameters between 0 and 2 times the initial value. The value of the parameters is depicted on the respective axis and the time to maximum Caspase 3 activation is colour coded from 0 (dark blue) to 20 hours (dark red). Right: Maximum Caspase 3 activation upon perturbation of two variables initial values. The maximum Caspase 3 activation ranges between 0 (dark blue) and 40 AU (dark red). The white indicates where the value exceeded the limit of the scale.

3.3.2.3 SASSy analysis indicates the sensitivity of the model depends on a large set of parameters

Using the program SASSy, sensitivity PCs and corresponding Singular values were calculated. The analysis shows that there are a high number of PCs with relatively high level of importance to the variance in parameter sensitivity (Figure 3.30). It also became clear by studying the first 3 PCs that even though they are high dimensional, it was a relatively small number of parameters that dominated each PC (Figure 3.31). Furthermore, a tail could be seen behind the top parameters propagating into lower PCs (Figure 3.32). Still, the combination of many PCs, even if they all were relatively low dimensional, added up to a large number of parameters with a significant effect on the sensitivity.

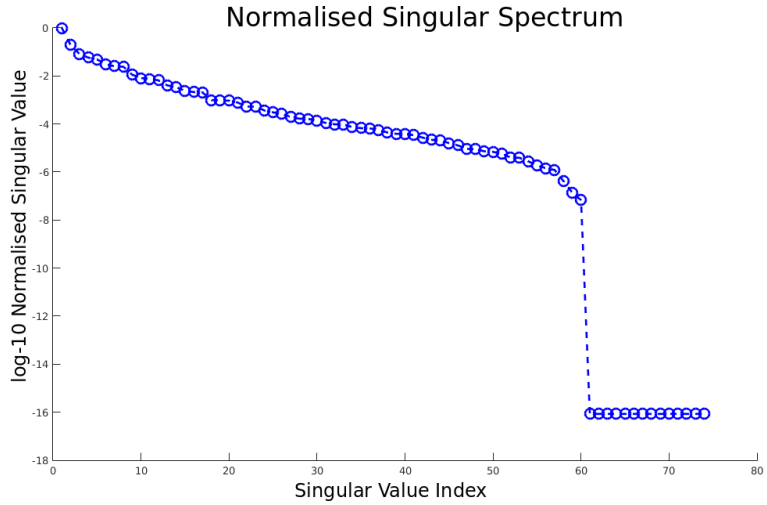


Figure 3.30: Sorted Singular values for SASSy analysis of larger apoptosis model. All Singular values are normalised so that the highest one is 1. Log_{10} Singular values on the y-axis.

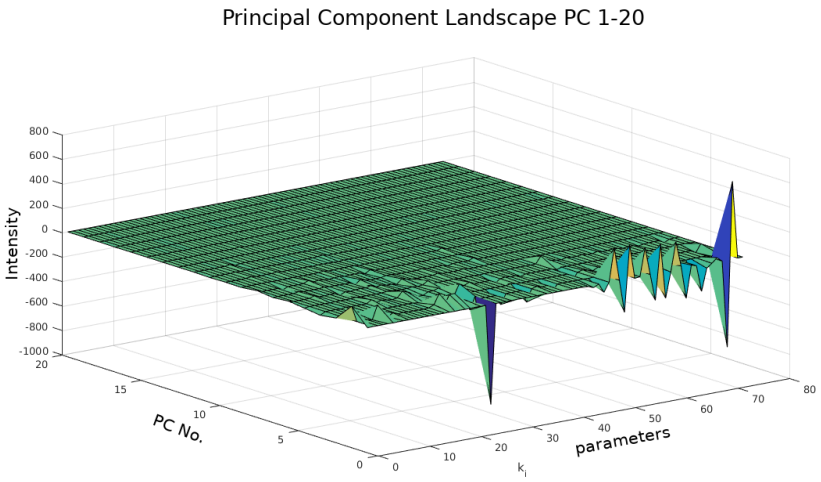
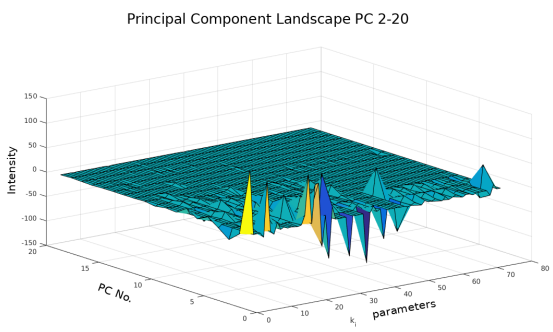
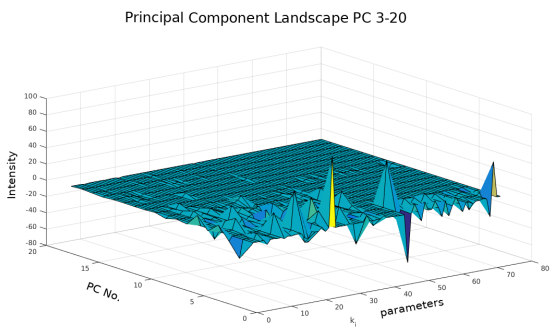


Figure 3.31: Principal component (PC) plot for the first 20 PCs in the SASSy analysis of the larger apoptosis model. Parameters on the k_i-axis with the first PC furthest to the front. Each PC has several parameters with a significant contribution.



(a) PC 2-20



(b) PC 3-20

Figure 3.32: PC plot of analysis using the SASSy program on the larger apoptosis model. (a) From second PC. (b) From third PC. Parameters on the k_i-axis with the first PC furthest down. Each PC has several parameters with is significant contribution.

3.3.2.4 SloppyCell analysis reveals two types of sensitivities

The model was run to time 20 hours and all parameters except actD, TNF, BHA, CHX, Tr, and $CytC_{free}$ were set to be optimisable. Around normal settings there was no other eigenvalue which was within 1% of the largest eigenvalue. This PC had a large number of parameters with a significant contribution to the sensitivity (Figure 3.33). The single parameter sensitivities, estimated by the Hessian with regard to single parameter perturbations spanned a much larger range of values compared to the analysis of the smaller model (Figure 3.34). This was true both when measuring all variables as experimental data and when measuring only the amount of activated Caspase 3.

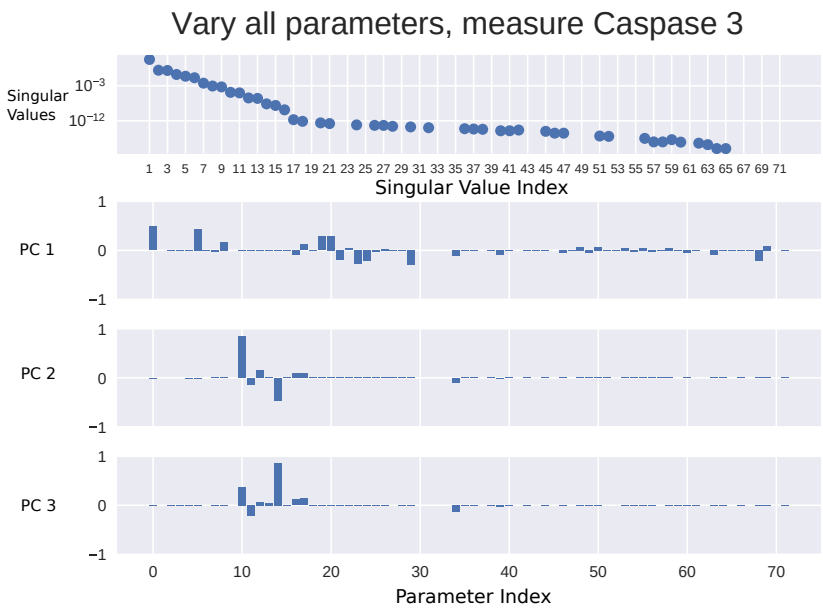
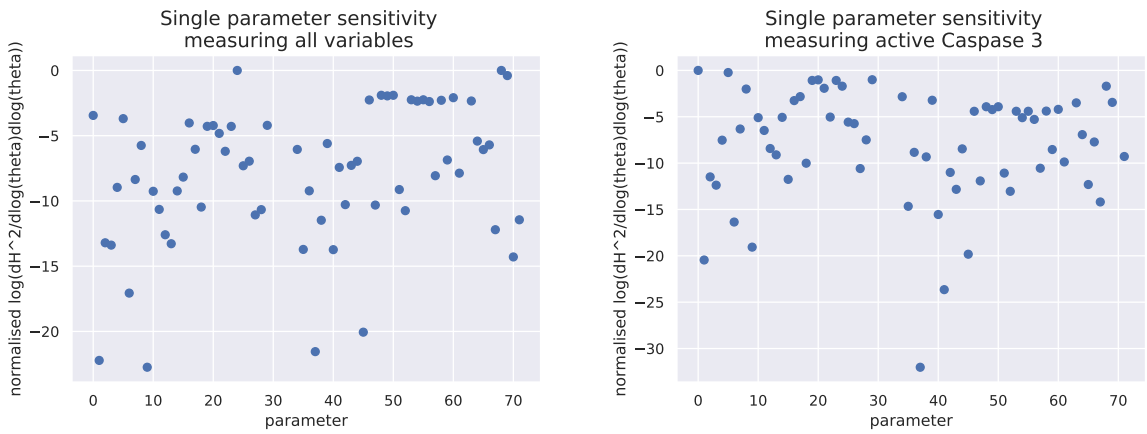


Figure 3.33: SloppyCell sensitivity analysis using normal variable and parameter setting and only active Caspase 3 as experimental data. From the top: eigenvalues, 1st PC and 2nd PC. Only one eigenvalue was within 1% of the largest eigenvalue namely the first one. Both the first and the second PC had contributions from a very large number of parameters.



(a) (b)

Figure 3.34: SloppyCell Hessian for single parameters ($d\chi^2/d\log\theta_j d\log\theta_j$) in the larger Apoptosis model. The Hessian spans several orders of magnitude both when measuring all variables (a) and when measuring only active Caspase 3 (b) indicating a large spread in sensitivities. Values have been normalised by the largest value. Small value indicates high sensitivity as only a small perturbation is needed for the constant model behaviour change.

3.4 Conclusion

Using variable and parameter scans, it was relatively simple to identify ranges of variable or parameter values where the system was more sensitive to further perturbations than in the normal cases. Furthermore, in the smaller apoptosis model these areas were very narrow, with an almost instantaneous flip from a short to a very long time to apoptosis. This was true for many variables in the larger apoptosis model as well, although perturbation of some variables had a more gradual effect on the time to apoptosis. However, it also quickly became clear that the sheer number of possible combinations of variables and parameters made it impractical to evaluate the effect a certain SNP could have in a biological system. In the larger apoptosis model, the analysis was restricted to initial protein concentrations resulting in scans of 16 variables. The analysis could have been extended to transcription and translation parameters as well, resulting in a 24-dimensional space. Even in the smaller system with only 19 parameters and 1 variable (not linked to any parameter) of interest the space of possible combinations of large perturbations was too large to fully examine. The analysis was therefore first limited to the parameters representing production rates. It was then further limited to evaluate how changes in one or two parameters affected the time to apoptosis and only how further perturbations in the already chosen parameters would affect the system. Even if only the resulting 2D-maps were to be considered and only the areas where further perturbations in the chosen parameters had a significant effect on the time to apoptosis, the analysis could have been expanded into all of the other parameters for those regions of each map.

The two sensitivity analysis methods, SASSy and SloppyCell, gave similar results, even though they measured slightly different things as, explained in section 3.1.2. SASSy indicated that the smaller apoptosis model was mainly controlled by a few parameters (Figure 3.18 and 3.19). The SloppyCell analysis on the other hand indicated a more complex sensitivity with many parameters contributing (Figure 3.20). Interestingly, SASSy indicated that the main contributors to the sensitivity were the production rate parameters of the four proteins involved in the model (Figure 3.19). Furthermore, this pattern was more or less unaltered when moving from an area where the time to apoptosis as simulated by the model was relatively insensitive (3,000 molecules of activated Caspase 8 at initiation), to an area where the model showed a much higher sensitivity (500 molecules of activated Caspase 8 at initiation). The pattern also remained regardless of whether the analysis was performed from time 0 to maximum concentration of activated Caspase 3, within a window around the maximum of activated Caspase 3 or from time 0 to 4000 minutes (Figure 3.18). This indicates that the sensitivity of the model is mostly decided

by a feature which is more or less the same in all cases, namely the short time frame when the activation of Caspase 3 and 8 overtakes the inhibition of IAP and BAR and the former see a rapid increase, whereas the latter decrease. This points to one of the problems with trying to use standard sensitivity methods to assess phenotype sensitivity, for which is it not designed.

In contrast to the relatively simple sensitivity of the smaller apoptosis model indicated by SASSY, the pattern for the larger apoptosis model was more complex. SASSy yielded a set of PCs with relatively slow decay in importance, where each of the PCs contained a smaller number of significant parameters compared to SloppyCell (Figure 3.31). SloppyCell on the other hand indicated that the sensitivity was mainly governed by the first PCs, which contained a large number of parameters (Figure 3.33). The end result was that, in order to explain a large part of the sensitivity, a large number of parameters would have to be considered. A further complication, hinted at earlier, is that most of the protein concentrations in the larger apoptosis model are set by initial variable values and the total concentration in various conformations, then remain constant throughout the modelling, or decay over time, in a fashion governed by a decay parameter, without being renewed. This means that it is not straightforward to measure the sensitivity of the model with regards to parameter changes and protein concentrations at the same time using these methods. Instead of creating PCs with both types of sensitivities in them, the model would have to be run several times, and parameter sensitivities, evaluated as the protein concentrations to be varied for each run. This would make it possible to get an idea of how the parameter sensitivities vary as variable values change, but not how they interact to affect the sensitivity of the model as a whole.

One constraint, which was not considered using SASSy, was which aspects of the system dynamics were of interest. Although the behaviour of the entire system upon perturbations might be of interest in some cases, in the cases studied here, it is arguably really only the dynamics of Caspase 3 which matters. As could be expected, whether data from all nodes were used or only that of active Caspase 3 had a big impact on the results of the analysis in the larger apoptosis model. Although the patterns of sensitivity were much more similar in the smaller apoptosis model, there was a slight difference both in the eigenvectors and the eigenvalues.

Although it could be argued that the activation of Caspase 3 is of far greater importance than the dynamics of the rest of the system, it is not necessarily the detailed dynamics of said component which is of interest, but the time to reach a peak corresponding to further activation of the pathway. Consequently, even if all of the other proteins

are ignored and only the sensitivity of Caspase 3 activation is considered, the results from SloppyCell or SASSy would not correspond directly to the sensitivity of interest.

To circumvent this problem, initially, the time frame in which the sensitivity was measured was adjusted in each case so that it would stop when Caspase 3 reached its peak activation levels. This allowed for the change in levels of activation to be used as a proxy for change in time, as shift in time would result in a shift in activity level during that time difference. However, this correlation between the shift in time and dynamics is not perfect as a perturbation could change not only the time of the peak but also the height. Furthermore, the correlation has some limitations with regards to the size and direction of the shift in time. If the time to the peak were to shift too far into the future upon perturbation, the increase in active Caspase 3 levels would not start within the set time frame and the correlation between the shift in time and dynamics would break.

The choice to focus on the four production rate parameters in the smaller apoptosis model was based on both the experimental results and the literature. These parameters were found to be the most sensitive in the SASSy analysis. Furthermore, most disease related SNPs are located outside of protein coding regions indicating that many of them have an effect on the expression of the protein rather than the function [30, 45]. This would mean that the four production parameters in the model would be the most likely to be subject to alteration by SNPs.

When limiting the analysis to these four parameters, there again were slight differences in the PCs when using different amounts of initial activation signal, suggesting different sensitivity patterns depending on the signal (Figure 3.23). This, however, only explores the possibility where the system is being sensitised due to a downgrade of upstream activation. As was shown by the parameter scan: the system could also alter behaviour, by mutations altering the parameters in the system.

From the parameter and variable scans two main areas of the phase space were identified, with either a high or a low maximum concentration of activated Caspase 3 in the larger apoptosis model, or a short and a long time to onset of apoptosis in the smaller apoptosis model. By closer examination, it became clear the two behaviours of slow and fast activation were separated by a curve in parameter space. The same could be said about the two behaviours of the larger apoptosis model in variable space. By extrapolation it is clear that this line extends into a surface in 3 dimension and a hyper-surface in higher dimensions. To fully explore the sensitivity of the model around this hyper-surface, one would have to analyse parameter combinations tracing all along the borders of states which did respond in time and the ones that did not. This is not only a daunting number of analyses, it would also be very difficult to compare the sensitivities between the states,

as a very small change in the initial parameters could have a large effect on the sensitivity spectrum obtained and it would be very difficult to be exactly the same distance from the separating border in every case. It was therefore concluded that the method would be able to give some insight into which SNPs would be most likely to cause a disease phenotype, given any specific set of mutations or other cellular events, bringing the system close to the border of the two phenotypes. However, it is not feasible to explore why one SNP is correlated with a disease and another is not in general, as that would require a weighted estimation of the sensitivity given all possible sensitivities for a system moving from a robust to a sensitised position in parameter space.

However, the parameter and variable scans, together with the parameter sensitivity analyses showed that, although the hyper space where dynamic behaviour changed rapidly was very complex, it was also very thin, meaning that in any one dimension, there was a very narrow window of parameter values with a higher sensitivity. This was especially true for higher initial amounts of activated Caspase 8 in the smaller apoptosis model. Considering how many genetic and epigenetic alterations could potentially alter the expression of a protein, and how they could interact to give a wide range of expression values, it is very unlikely that a system would, by chance, end up precisely in the parameter value window where the system is sensitised.

Considering how thin this hyper space of increased sensitivity is, it is as if it is acting like border separating the two behaviours. If the problem is viewed from that perspective, the question would then not be, which SNP has the largest effect once the system is sensitised, but how likely is a SNP to push the system over the border, given a random set of mutations or other alterations, acquired during the course of a persons life. This will be the scope of the coming chapters of this thesis.

Chapter 4

Separatrix Analysis

4.1 Theory

In chapter 3 the limitations of standard sensitivity analysis tools were exposed, when trying to use them to examine the sensitivity of phenotype change. Using the smaller and larger apoptosis model by Eissing *et al.* and Schlatter *et al.* respectively it was seen that for large parts of parameter or variable-state space there was very little sensitivity in the part of the model dynamics interpreted as representing phenotype change. Furthermore, as the initial settings moved in state space, representing accumulations of mutations over the course of an individual's life, there was a very small window of opportunity where the system exhibited a heightened sensitivity in the important dynamics. In fact, it seemed almost as if the two phenotypes, responsive and non-responsive, were separated by a surface and crossing this surface resulted in near-instantaneous switching.

In this chapter, a method will be developed and explored, making use of this surface in order to assess phenotype sensitivity. First, the suitability of this surface as a measure for phenotype sensitivity will be explored conceptually. Then a method for using the surface will be presented and finally the new method will be explored in principle and using simulated Single Nucleotide Polymorphism (SNP) data.

4.1.1 Separatrix surface

Considering a general dynamical system, if the output of the system depends on one parameter or the initial state of one variable (from here on only parameters will be mentioned, however the concepts discussed also apply to initial variable values), then as that parameter is being perturbed away from its initial value, the output will start to deviate. If enough is known about the phenomenon said system is modelling, then there will be a point along the line of possible values, where one can consider the output no longer being

within the range of normal system behaviour. In biological systems that might correspond to an apoptosis network yielding a time to activation which is no longer consistent with normal cell behaviour, i.e. effectively no apoptosis. Alternatively, it might be a cell cycle model which no longer responds to control signals in a way characteristic of normal cells. This point along the parameter value line can be seen as separating two sets of values yielding distinct model behaviours and will from here on be referred to as a separatrix point.

If the system were to have two parameters affecting the output, combinations of those parameters would form a separatrix curve in a two dimensional parameter space separating the two sets of behaviours. Additionally, a system with three parameters would form a surface and systems with more parameters would form a hypersurface in a hyperspace.

Note that the use of the term separatrix here is different from that traditionally used in mathematics. In mathematics a separatrix is traditionally defined as the separation between basins of distinct dynamical behaviour, such as the rotation and oscillation of a pendulum or the two steady states of a bistable switch. In this thesis, the separatrix will be defined as separating biological behaviour. Since biological behaviour is not always clearly defined and not always easy to relate to dynamical behaviour of a model, these will have to be defined for each phenotype and model explored.

If the separatrix defines the border between cancer and non-cancer phenotypes, then the risk of acquiring cancer could then be seen as the risk of crossing over from one side of the separatrix to the other. In the case of apoptosis that would mean shifting from a position of "normal apoptotic response" to a position of "no apoptotic response" such as might initiate a tumour. A change in any parameter accounted for in the separatrix, or any combination thereof, could potentially cause the system to cross over. The sensitivity of the system with regards to that parameter could be seen as the degree to which a given change in that parameter changes the likelihood of the system to cross over.

There are several ways in which the notion of a separatrix could be used to evaluate phenotype sensitivity. In this thesis the suitability of the average distance between starting point and separatrix as a general measure of phenotype sensitivity will be assessed.

This approach relies on a number of assumptions:

1. There exists a clear definition for separating the behaviour of the system into a normal and an abnormal phenotype.

2. Over any given time-frame, cells in the body would accumulate mutations and epigenetic changes which would shift the position in parameter space.

3. The likelihood of a set of mutations to move the system to a certain position in parameter space can be related to the distance between the new location and the start location.

Before evaluating the suitability of the distance to the separatrix as a measure for phenotype sensitivity, one would first have to decide what is meant by the distance between a point and a surface, and further, how that relates to the biological question in mind. In general the distance from a point to a surface or a volume is taken to be the shortest distance between the point and any point on the surface, within the volume. For many purposes that is a reasonable definition.

However, if this concept is being applied to the likelihood of a biological system crossing over a separatrix, it is not at all clear that the point closest to the starting point is a good approximation for the surface. This point would correspond to the least amount of perturbation needed to reach the surface, but if the parameter space is multidimensional, this could require larger or smaller perturbations in all or, most of the parameters. If the likelihood of a perturbation of a parameter is small to begin with, it is not at all clear that two, three or four small to medium range perturbations are more likely to take place than one large one. If the system can take a large number of paths to cross the separatrix, it seems more suitable that the distance to the separatrix should be considered a function of all the possibilities of crossing over.

In a system with one parameter affecting the outcome, regardless of the distribution of possible values the parameter can take given the possible genetic alterations and their individual likelihood, the likelihood of crossing the separatrix point is proportional to the ratio of the density of the distribution on the other side of the separatrix to the entire distribution (It is assumed that, given physical limitations, the distribution has an finite upper limit).

By extension, if the shape of the distribution is independent of where in parameter space the system is starting, then an instance closer to the separatrix must have a higher chance of crossing over to the other side than an instance starting off further away. This is because an instance closer to the separatrix will extend the part of the distribution on the other side of the separatrix. Hence, the likelihood of crossing over the separatrix is relatable to the distance between the initial point and the separatrix. Furthermore, if the likelihood of crossing over in any one direction is proportional to the distance between the starting point and the separatrix point in that direction, then the likelihood of crossing over could be seen as the average of the distances to every part of the surface.

If the system is already sensitive in one or several directions at the original location of the parameter space, then such sensitivities could be detected by standard sensitivity

methods as those discussed in Section 3.1.3. However, in the case of disease associated SNPs, it is understood that the system is generally far away from the separatrix. If the analysis is limited to parameters for which this is true, the likelihood of crossing over the separatrix could be seen as the average distance to all points on the separatrix at a comparable distance from the initial location. This puts a practical limitation on the method, as the parameters included in the analysis cannot have too wide a variation in sensitivity, which is something to keep in mind when designing the experiment.

4.1.1.1 Specific limitations for this thesis

The presence of something which could be interpreted as a separatrix surface was seen in Chapter 3, using the smaller and larger apoptosis model as study examples. For the smaller apoptosis model, the initial amount of activated Caspase 8 acts as an initial activator from an upstream pathway and the following burst in Caspase 3 activation is interpreted as onset of apoptosis. When decreasing the amount of initial activated Caspase 8, the time it took for the system to commit to apoptosis increased. If the amount of initial activator was low enough, no burst in activated Caspase 3 (onset of apoptosis) could be seen, even with the large time window set. The point at which any further decrease in initial activated Caspase 8 would cause the system to not respond at all can be seen as the separatrix point. It was also shown that the same behaviour could be seen when altering the parameters guarding protein translation in the system, and also when altering two parameters at the same time. In both cases a separatrix line could be identified. This was shown to be true for all four model parameters and, by extension, alteration of all four parameters at the same time would yield a four-dimensional separatrix surface. Similar results were shown for pairs of initial variable values for the components in the larger apoptosis model, and following the same logic the concept could be extended into a multidimensional separatrix surface.

Although the time to activation of apoptosis changed as the system got closer to the separatrix, for most of the parameter or variable space, this change was not very large and as the system got close to the surface there was a rapid shift towards much longer response times (i.e. longer time between initial signal and onset of apoptosis). In any case, from a biological point of view, small changes in response time do not have much effect on the system, since even if a cell managed to escape apoptosis just before mitosis, the daughter cells would have time to respond. Similarly, even large changes in response time are of little importance if the system still has time to respond, as they would have the time of a full cell cycle. With that in mind, the separatrix surface, if time to apoptosis is

set large enough, can be seen as a reasonable, although very simplified way of classifying sets of cells which could create cancers and cells which could not.

To ensure that all points on the separatrix are of somewhat equal importance, the surfaces used in this chapter will be confined between 0 and 2 times the initial values of each parameter or variable. The problem of including points far beyond these limits can be illustrated by considering a system with the following criteria (Figure 4.1):

1. the initial point is located between 0 and parts of the separatrix along the y-axis,
2. parts of the separatrix are located between 0 and the initial point along the x-axis,
3. points on the separatrix far beyond 2 times the initial value are taken into account (Figure 4.1, blue area).

In this case there would be an asymmetrical distribution of points around the initial value. The mean distance would then be dominated by perturbations along the x-axis where the separatrix extends beyond 2. Even if a perturbation along the extended axis were to move the system closer to some points, this decrease in distance would be cancelled by the increased distance caused by moving further away from all the points in the extended region.

In the case of genetic and epigenetic effects on a biological system this limitation is not as severe as might first be thought. The lower limit of any parameter is naturally zero, corresponding to a complete knock-out of a gene or a mutation that destroys its ability to perform its function. Although the upper limit might not actually be 2 in reality, there is usually a limit for how much a gene can be expressed before it either becomes toxic to the cell or starts to interrupt other essential functions by high-jacking the translational system. Mutations affecting function, rather than expression rarely render the protein more efficient than it already was and 2 might then seem like an over-estimation. There are several mutations which are known to render kinases constitutively active. However, these mutations would represent structural differences in the model, as opposed to shifts in parameter values. For that reason, these types of mutations will not be considered in this thesis.

In this chapter, the separatrix will be restricted to parameters of the smaller apoptosis model which govern protein production or variables of protein concentrations in the larger apoptosis model. This decision is based on two facts. First, it was shown in Chapter 3 that the production parameters were among the most sensitive in the smaller model, with regards to phenotype sensitivity. Secondly, although the exact function of most SNPs

is not known, eQTL data and expression profiles indicate that many of them do indeed affect transcription levels.

Furthermore, by limiting the analysis to expression altering parameters, experimental data could later be used to alter the models and estimate outcome. This gives the possibility to assess the suitability of the method on experimental biological data.

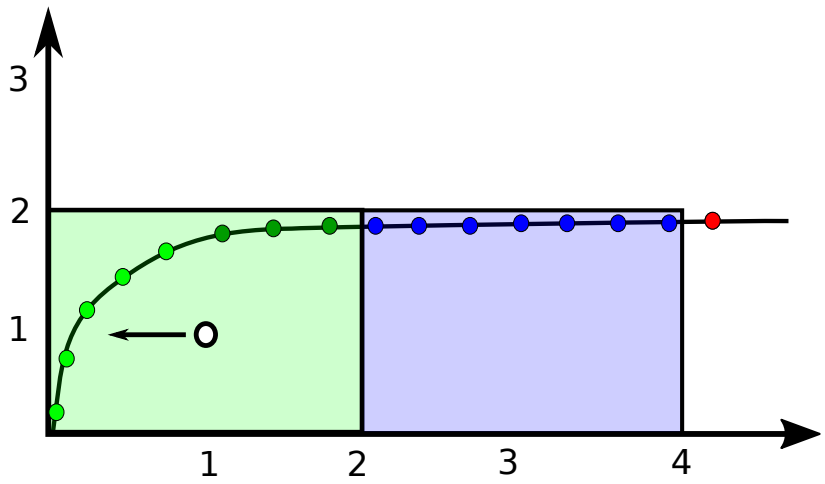


Figure 4.1: Illustration of the problem of extending parameter space beyond 2 times the initial value. A Perturbation along the x-axis would bring the system closer to a set of points between 0 and 1 on that axis (bright green), but further away from a set of points above 1 (dark green). As the space extends further away from the starting position there will be ever more points which the system will be further away from (blue). If the space is extended long enough the average distance from initial point to points on the separatrix surface (as defined in Section 4.1.1) will increase as the system moves closer to the basin of abnormal behaviour (decreased x). Additionally, as was pointed out in Section 4.1.1, if the sensitivity of one of the parameters is much higher than the sensitivity of the other parameters (so that the separatrix is much closer in one direction than in another), then a similar problem of unbalanced distribution of points arises, even if the surface is limited between 0 and 2 times the initial values.

4.2 Materials and Methods

4.2.1 Models

4.2.1.1 Smaller apoptosis model

The small apoptosis model published by Eissing *et al.* (2004) [95], introduced in Section 3.2.1.1 and used throughout Chapter 3 was further used in the work of this chapter. The

four parameters corresponding to production rate of the four proteins Caspase 3, Caspase 8, IAP and BAR were targeted for further analysis. The initial activated Caspase 8 signal was set to either 3,000 or 10,000 molecules and was subtracted from the total amount of inactive Caspase 8 before starting the simulation. All equations and parameter settings can be found in Appendix B.

The model was run up to 5,000 minutes after initial activation signalling. If no sufficient Caspase 3 peak could be measured within this time, the system was classified as a non-responder.

When parameters were perturbed from the standard values, the system was first run to 5000 minutes with initial Caspase 8 activation set to zero, to allow for the system to find its new steady state. The final steady state concentrations were then used as initial concentrations in the actual run, with initial activation of Caspase 8.

4.2.1.2 Larger apoptosis model

A model of TNFα induced apoptosis signalling published by Schlatter *et al.*, (2011) [91], implemented in 3.2.1.2 was further used in this chapter. The node corresponding to translational inhibition by cycloheximide was always set to 0, as was the node corresponding to the antioxidant butylhydroxyanisol (BHA) and node corresponding to translational inhibition by actinomycin D. The activation signal coming from TNF was set to 100 and FasL was set to change from 0 to 100 after 12 hours, as described in the paper. All equations and parameter settings can be found in Appendix C. The onset of free Reactive Oxygen Species (ROS) production was modelled by the function:

$$ROS_{free}(t) = \frac{1}{0.03 \times 2\pi} e^{\frac{1}{2}(\frac{t-4}{0.03})^2} \times 100 \times (1 - BHA) \tag{4.1}$$

resulting in a short burst of 100 ROS units being released after 4 hours. The model was run for 20 hours unless stated otherwise.

The analysis was performed on two sets of variables: all the variables which were not zero at time zero and with nodes that were not governed by production parameters; all variables at the core of the model (Caspase 8, Caspase 3, XIAP, Bim, Bid, Bcl2 and BaxBak), corresponding to the function of the smaller apoptosis model.

Limiting the separatrix space of the larger model to the pathway equivalent of that covered by the smaller model allowed for a better comparison of the results between the models and an evaluation of how the results translate from a smaller to a larger model.

4.2.2 Phenotype sensitivity analysis tool

4.2.2.1 Defining separatrix surface and measuring distance from starting point

To define the separatrix, points of the separatrix surface were determined using an adapted method previously published by Cavoretto *et al.* [103, 104]. The way the full method works is outlined in Figure 4.2 and Algorithm 2 and 3. For each parameter under investigation, a range of starting values were chosen and arranged so that they formed a Latin Hypercube (using pyDOE v.0.3.8). For each set of parameter values, a point on the surface was searched for along the axis of each parameter value. For each parameter consecutively, the model was run with the maximum and minimum of the search range. If the two runs gave results indicating they were on different sides of the separatrix surface, the search was continued through a binary search algorithm with a predefined number of iterations (the number of iterations varied between 5 and 20). For each iteration, the middle point between the maximum and the minimum was identified. The model was run with the minimum and middle parameter values. If the two runs gave results indicating they were on different sides of the separatrix surface, the middle point was taken as the new maximum and the search was iterated. If both runs were on the same side of the separatrix, the middle point was taken as the new minimum and the search iterated. For the last iteration, either the middle or the maximum was taken to be the last approximation of the point on the surface.

To calculate the distance between the system starting point and the separatrix surface, the length of each vector between the starting point and the points identified to be close to the surface was calculated and a mean of them taken to represent the mean distance. To evaluate the sensitivity of a parameter or variable, the starting position of each parameter or variable was perturbed individually and the distance to the surface was calculated again.

4.2.2.2 GWAS simulator

A python script was written, simulating SNPs linked to parameters or initial concentrations of nodes in an ODE based biological pathway model and estimating how each SNP affected the risk of a system to cross the separatrix given accumulation of mutations over the course of a lifetime. In short the script functions as follows: A SNP is created having a random target parameter and an effect k on that parameter, randomly drawn from a normal distribution centred around 1.0 (σ=0.05), as well as an allele frequency randomly chosen from a uniform distribution between 0.0 and 0.5.

Algorithm 2 Finding points on the separatrix surface

```
search_range = max and min of the space explored
iterations = the number of iterations to narrow down location of separatrix point
for all indices p in list of parameters in analysis do
    for all vectors v of initial parameter values do
        start and end = v
        start[p] = minimum of search_range  ▷ Position of parameter p in vector start
        end[p] = maximum of search_range
        if SeparatrixTest(start,end) == True then    ▷ start and end give rise to different phenotypes
            SearchPoint(start, end, iterations)    ▷ Hone in on the actual separatrix point
        end if
    end for
end for

function SearchPoint(start, end, iterations)
    middle = start
    for iteration in iterations do
        middle[p] = (start[p] + end[p])/2
        if SeparatrixTest(start,middle) == True then
            end = middle
        else
            start = middle
        end if
    end for
    return end                          ▷ Return the upper vector as separatrix point
end function

function SeparatrixTest(start, end)
    peak1 = maximum of activated Caspase 3 (using start)
    peak2 = maximum of activated Caspase 3 (using end)
    if peak1 < threshold and peak2 > threshold then
        return True                     ▷ Separatrix point is within interval
    else if peak1 > threshold and peak2 < threshold then
        return True                     ▷ Separatrix point is within interval
    else
        return False                    ▷ Separatrix point is not within interval
    end if
end function
```

Algorithm 3 Calculating the mean distance from starting point to the separatrix surface

```
function DistanceToSurface(separatrixPoints, startPoint)
    lengthList = empty list
    for all points in separatrixPoints do
        append (point - startPoint) to lengthList
    end for
    return mean of lengthList
end function
```

An individual is then simulated with genotypes for each SNP chosen from binomial distributions with likelihoods according to predefined allele frequencies. Each parameter is then perturbed $\prod \varepsilon_i^{x_i}$ times the normal value (as described in the original publication of the model being studied), where ε_i is the effect of the SNP$_i$ and x_i is the number of risk alleles for that SNP. The individual's response to apoptotic signal is then simulated as described in Section 3.2.1.1 and if the response time is within normal range (5,000 minutes for the smaller apoptosis model and 25 hours for the larger apoptosis model), the individual is kept. Otherwise it is discarded and a new individual is simulated. This ensures that, although every simulated individual would have different start point in parameter space due to their individual genotypes, all simulated individuals would have a life compatible phenotype at "birth". A second perturbation is then applied to each parameter under investigation, drawn from a separate distribution (for the smaller model: type gamma with k=2.2 and θ=1, for the larger model: type log normal with mean=0 and σ=0.5) simulating a life time accumulation of mutations. The parameter values were chosen so that the distributions would resemble those of RNA expression values in cancer cells (a heavy centre around 1 and a long, thin tail). The response is once again simulated, and the individual characterised as having a cancer phenotype if it does not respond within the time frame of the simulation. If it does respond, the individual is characterised as healthy. When a sufficient amount of individuals in both healthy and cancer groups has been collected, Odds ratios for cancer are calculated for all SNPs in the simulations, according to standard methods.

$$ODDs - Ratio : \frac{D(A)/D(B)}{H(A)/H(B)} \tag{4.2}$$

where:

A and B are the two alleles

H(x) and D(x) are the sum of healthy and diseased individuals with allele x, respectively.

4.2.2.3 Estimating the correlation between simulated SNPs and model characteristics

Using 50 simulated SNPs, linked to parameters in the model, subsets of SNPs, ranging in number from 10 to 50, were chosen and individuals were simulated with defined genotypes, according to the method described in Section 4.2.2.2. For each individual the effect of the genotype on the model parameters was taken into account and the model was run. Then the individual's risk score:

$$\sum(\mid \log_e(\text{Odds ratio})_i \mid \times x_i) \tag{4.3}$$

where:

x_i is the number of risk alleles for SNP_i over all risk associated SNPs,

was calculated. Finally the correlation between the risk score and either time to apoptosis or distance to separatrix was tested using a linear least-square regression method. There is no clear consensus in the fields as to how disease risk or genetic effects on model parameters should be modelled. However, a log-additive model seems to be the most common way of modelling SNPs effect on risk of developing disease and consequently, it was chosen for this work. A multiplicative model seemed most reasonable when considering the genetic effects on the models most likely to be of concern in this work (change of production rate) and was chosen for modelling the effect of SNPs on the model parameters. To assess the robustness of the method, the analysis was repeated several times, with different numbers of SNPs and different sizes of risk scores.

If the method were to be applied to real biological data, there would be errors and noise in every step of the analysis, making it more difficult to identify any links between the risk of developing cancer associated with a certain genotype and the parameter perturbation that genotype is causing. To investigate how noise in the data set would affect the likelihood of finding a significant regression the analysis was also performed, linking only a random subset of the SNPs to the parameters, while all of them contributed to the risk score.

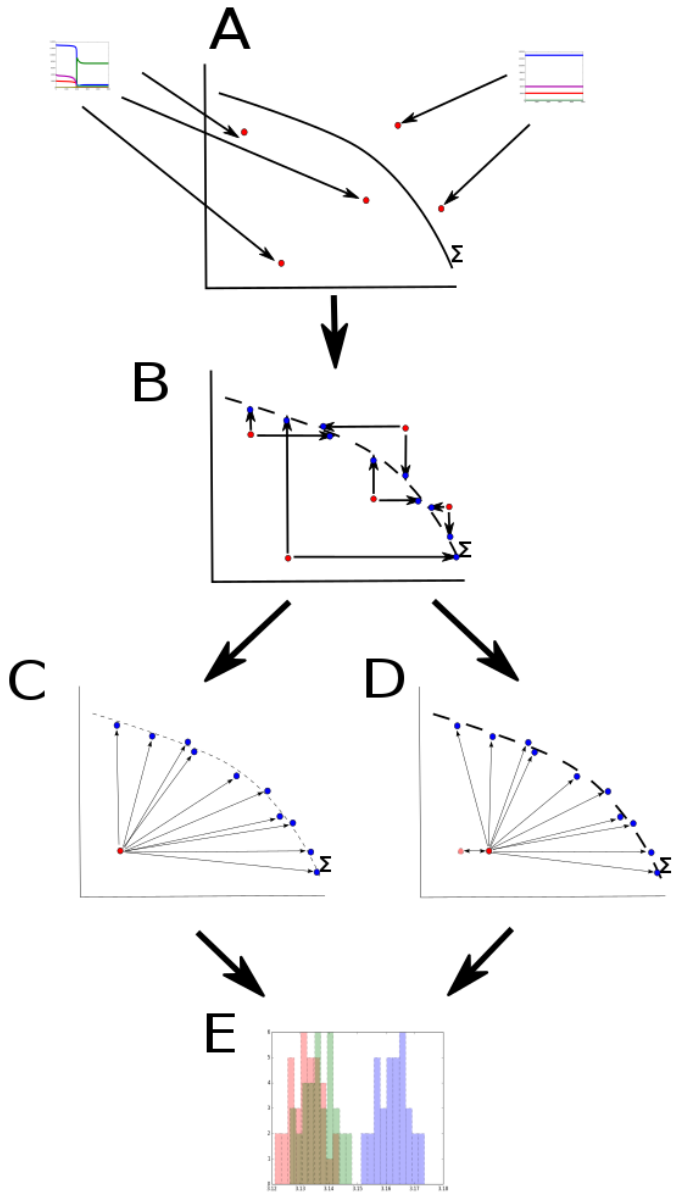


Figure 4.2: Illustration of method for generating separatrix surface and calculating distance from the starting point to the separatrix surface. A: points are randomly chosen in parameter-state space in a way which tries to cover as large a part of the space as possible. B: along each axis a point on the separatrix surface is searched for by looking for increasingly small intervals in which the phenotype switches. C: the average distance is calculated between the starting point and the points on the surface. D: the starting point is moved, corresponding to a perturbation caused by a SNP and the average distance is calculated again. E: the differences between the two measures can be compared over several surfaces or individuals.

4.3 Results

4.3.1 Mean distance to separatrix surfaces quickly converge and precision increases as number of points and number of search cycles increase

In this work a numerical method was developed for finding points on the separatrix and relate them to the distance between a starting point and the separatrix as a whole. When creating a separatrix surface using this method, two parameters have to be set: the number of points on the surface and the number of cycles used to narrow down the interval within which the point is situated (see Algorithm 2). To investigate how much of an effect these parameters had on the shape of the generated surface and downstream analysis several surfaces were generate using varying number of points and cycles (iterations in SearchPoint() in Algorithm 2). For each surface the mean distances from starting point (1 in all dimensions) to surface was measured. Using the smaller apoptosis model by Eissing *et al.*, when increasing the number of points on the separatrix surface, the change in mean distance from the starting point very quickly approached zero (Figure 4.3 and 4.4). The mean distance as well as the deviation between runs showed much less dependence on the number of iterations for narrowing down the position of the separatrix points than the number of points used. This was especially true at the upper range of number of points chosen.

Using the larger apoptosis model, a similar trend could be seen, where the variation between simulations quickly decreased as the number of points on the surface increased. However, whereas the number of points used had little effect on the mean distance, the precision (i.e. the variation in mean distance between surfaces) never did converge within the range of points explored in this analysis, but continued to decrease, although at a slower pace. This was true both when the surface was containing all constant concentration variables in the model and when confining the analysis to the 6 variables at the core of the model, corresponding to the smaller apoptosis model (Figure 4.5 and 4.6).

In summary, when generating the separatrix surfaces, the number of points initially used to find the points of the surface proved to have a much larger impact on the precision of the surface than the number of cycles used to narrow down on the interval within which the surface was located for each point.

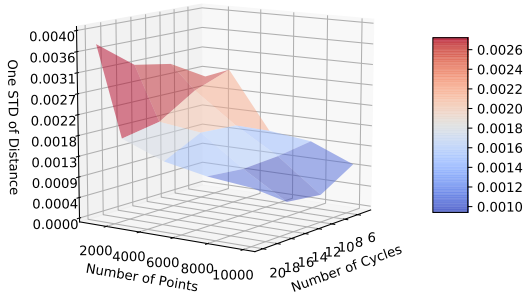


Figure 4.3: One standard deviation of distances calculated over 10 surfaces generated using different latin hypercubes for the smaller apoptosis model, using 10,000 molecules as initial activated Caspase 8. Along the x-axis the amount of hypercube points which are being fitted to each axis in parameter space is being varied between 1,000 and 10,000. Along the y-axis the number of times the range within which the separatrix point is located is being halved, is varied from 5 to 20. The spread of the distances is quickly decreasing as the amount of points are increasing, but the rate with which it is decreasing decreases as the number of points increases. The number of cycles (iterations in SearchPoint() in Algorithm 2) of narrowing down on the separatrix point did not prove to have as much of an affect on the variation of distances calculated between surfaces.

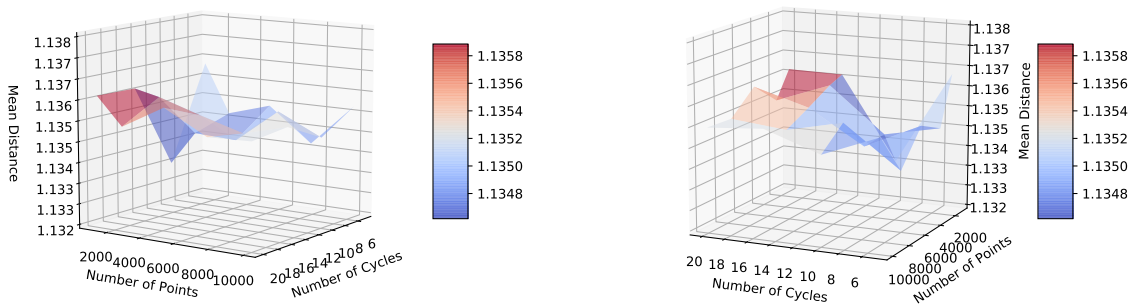


Figure 4.4: Mean of distances to separatrix calculated over 10 surfaces for the smaller apoptosis model, using 10,000 molecules as initial activated Caspase 8. Both plots show the same data from different angles. Along the x-axis the amount of hypercube points which are being fitted to each axis in parameter space is being varied between 1,000 and 10,000. Along the y-axis the number of times the range within which the separatrix point is located is being halved, is varied from 5 to 20. With the exception of very low number of points, there was not much difference between the setting. The system seemed to settle around a narrow range of distances using between 5,000 and 10,000 points and between 15 and 20 cycles (iterations in SearchPoint() in Algorithm 2) for narrowing down on the surface points.

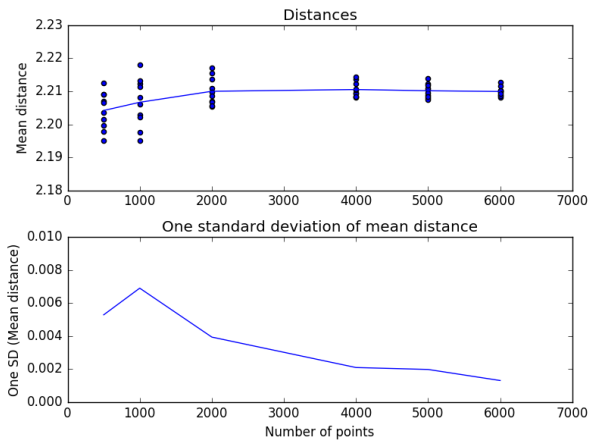


Figure 4.5: Top: Mean of distances calculated over 10 surfaces for the larger apoptosis model, using 10 cycles (iterations in SearchPoint() in Algorithm 2) of narrowing down on the separatrix surface. Along the x-axis the amount of hypercube points which are being fitted to each axis is being varied between 500 and 6,000. Bottom: One standard deviation of the mean distances.

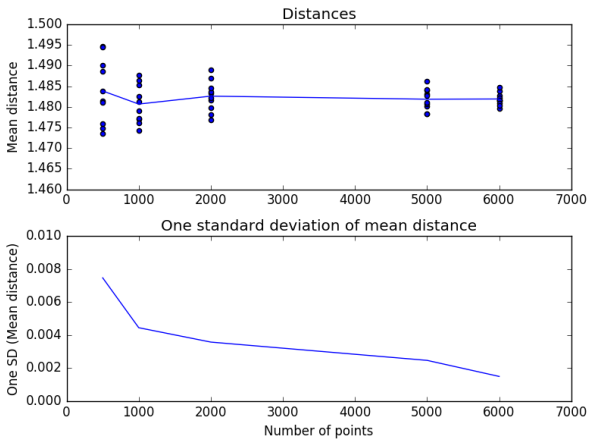


Figure 4.6: Top: Mean of distances calculated over 10 surfaces using a smaller set of variables in the larger apoptosis model, using 10 cycles (iterations in SearchPoint() in Algorithm 2) of narrowing down on the separatrix surface. Along the x-axis the amount of hypercube points which are being fitted to each axis is being varied between 500 and 6,000. Bottom: One standard deviation of the mean distances.

4.3.2 Phenotype sensitivity analysis

4.3.2.1 Perturbations in the smaller apoptosis model result in clear differences in distance to the separatrix

When perturbing each parameter corresponding to the production rates of the four proteins in the model individually, the distance between the new starting point and the separatrix surface changes in an expected fashion. A decrease in either of the two inhibitors, IAP or BAR, brings the system further away from the separatrix, whereas a decrease in either of the activators, CASP3 or CASP8, brings the system closer. This behaviour is in agreement with what would be expected, since a decrease in the amount of inhibitor would decrease the time to apoptosis and render the system less "cancer-like". On the other hand, a decrease in an activator would increase the time to apoptosis and render the system more "cancer-like". When comparing the actual distances after small perturbations on ten different surfaces to the distances to the surfaces without any perturbations, there was a clear overlap of the distributions of distances for the ten surfaces before and after the perturbation (Figure. 4.7), meaning that the variance in distance between different calculations of the surface was generally larger than the effect a perturbation had on the distance to each surface. However, when taking the same surfaces and comparing the change in distance for each generated surface (i.e the difference in distance to the same surface with and without a perturbation of the initial point) the results were very consistent (Figure. 4.11a), indicating that although there was a difference between the generated surfaces, each of them were capable of representing the actual surface. Furthermore, the direction of the change in distance depends both on the initial concentration of activated Caspase 8, which dictates how far away from the starting point the separatrix surface is (As can be seen in Chapter 3), as well as the amount and direction of the perturbation. For example, when initiating with 10,000 molecules of activated Caspase 8, upon perturbations of IAP and BAR the distance changes in the same direction, for any perturbation between -0.1 and 0.05 times the initial value, as does the distance upon perturbations of Caspase 3 and Caspase 8 (Figure 4.8). When the amount of an inhibitor (IAP and BAR) is increased, the system moves closer to the separatrix, i.e. it becomes more likely to stop responding to an apoptosis signal. Likewise, when the amount of an activator (CASP3 and CASP8) increase the system moves further away from the surface, making it less likely to stop reacting to activation signal.

However, if the system is initiated with 3,000 molecules, small perturbations of IAP cause the distance to change in the opposite direction of the distance resulting from perturbation of BAR, and in the same direction as that resulting from perturbations

of either Caspase 3 or Caspase 8 (i.e. decreased inhibitor decreases likelihood of not responding.) (Figure 4.9). If the perturbation is large enough in the negative direction, however, the shift in distance changes direction and moves with that of BAR, until it shows the expected behaviour again, i.e. opposite behaviour of perturbations of either Caspase 3 or Caspase 8 (Figure 4.10). The results from the distribution of distances were consistent with that of changes is mean distance of each surface (Figure 4.11b).

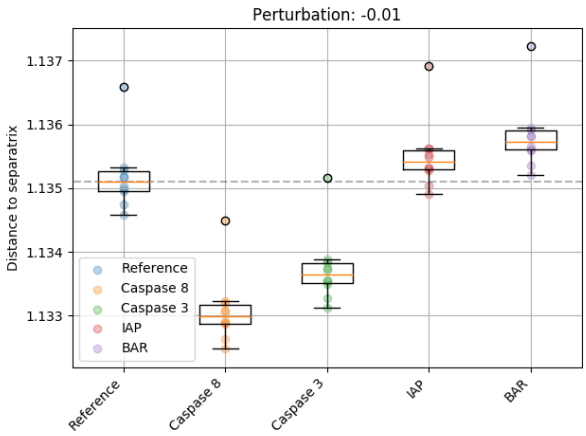


Figure 4.7: Distance from starting point to separatrix surface, calculated as mean distance to all points on the surface. The surface is calculated with 10,000 molecules of initial activated Caspase 8 and each parameter is perturbed -0.01 times the initial value. Blue data indicate distances from original starting point to 10 surfaces whereas other data indicate distances after perturbation of the production rate parameter of the corresponding protein. As the parameters corresponding to Caspase 3 and Caspase 8 production rates are decreased the distance to the surface decreases as well, whereas a decrease in the parameters corresponding to production parameters for IAP and BAR result in an increase of the distance.

A closer inspection of the distribution of the separatrix points along each axis, shows that the surface is skewed towards smaller values of IAP when initiating with 3,000 molecules, compared to when initiating with 10,000 molecules, (Figure 4.12 - 4.15). this indicates that there are more possibilities for the system to cross over from a responder state into a non-responder state, when IAP is down-regulated compared to when it is up-regulated, even though it is acting as an inhibitor. The same skewness could be seen along the axis of Caspase 3 and Caspase 8 in both scenarios, indicating that there are more possibilities to cross over into a non-responding state with lower levels of activators, regardless of the strength of initial activation signal, as would be expected. The distribution along the axis of BAR was much more uniform in both cases.

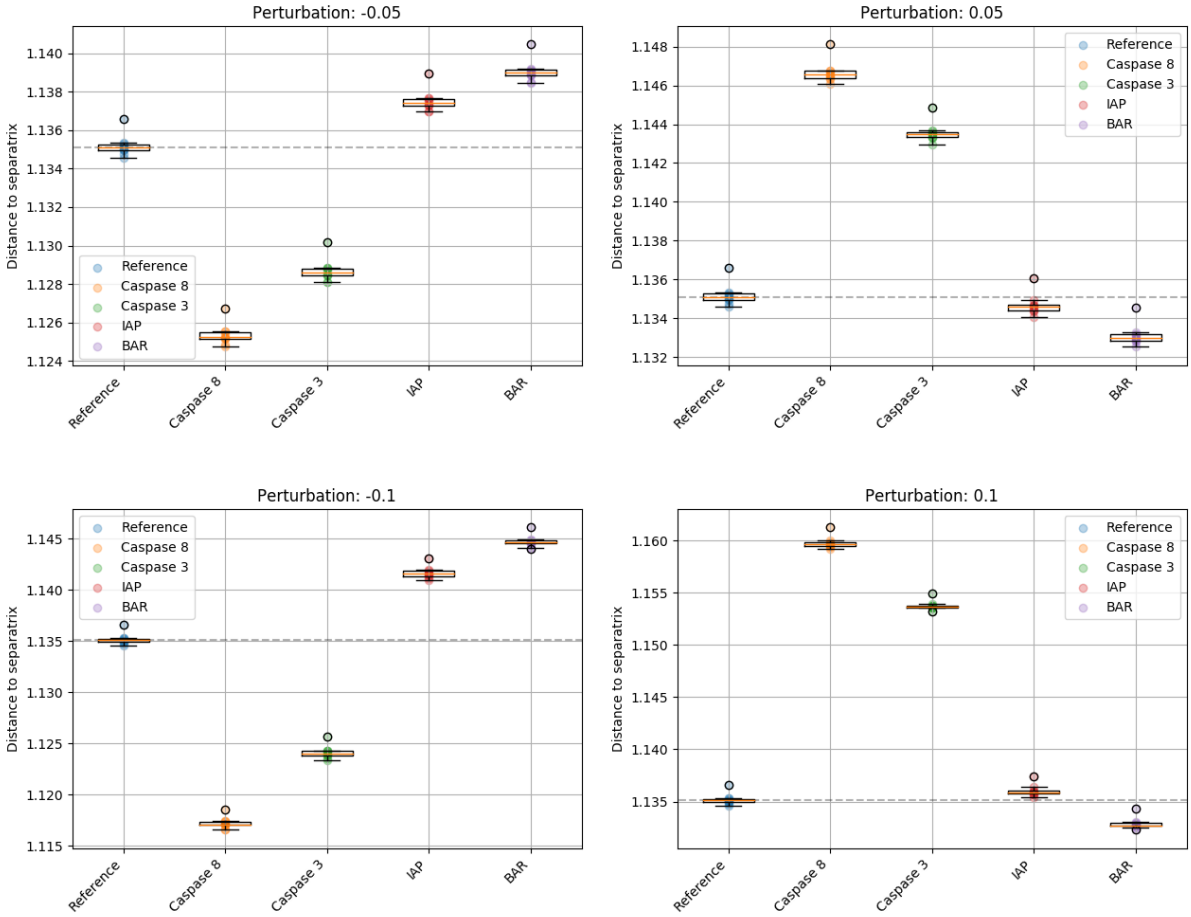


Figure 4.8: Distance from starting point to separatrix surface, calculated as mean distance to all points on the surface. The surface is calculated with 10,000 molecules of initial activated Caspase 8 and each parameter is perturbed: top left; -0.05, top right; 0.05, bottom left; -0.1, bottom right; 0.1 times the initial value. Blue data indicate distances from original starting point to 10 surfaces whereas the rest indicate distances after perturbations of respective production rate parameter.

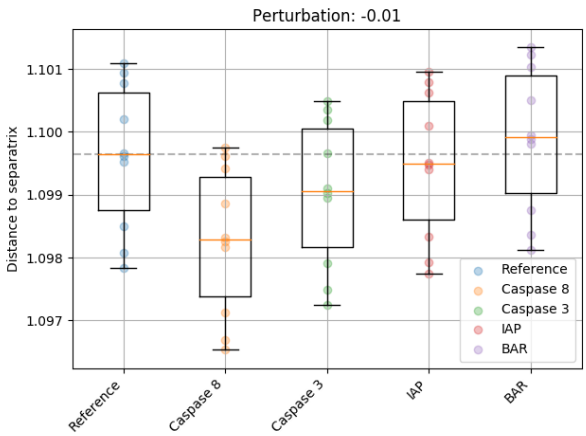


Figure 4.9: Distance from starting point to separatrix surface, calculated as mean distance to all points on the surface. The surface is calculated with 3,000 molecules of initial activated Caspase 8 and each parameter is perturbed -0.01 times the initial value. Blue data indicate distances from original starting point to 10 surfaces whereas the rest indicate distances after perturbation of respective production rate parameter.

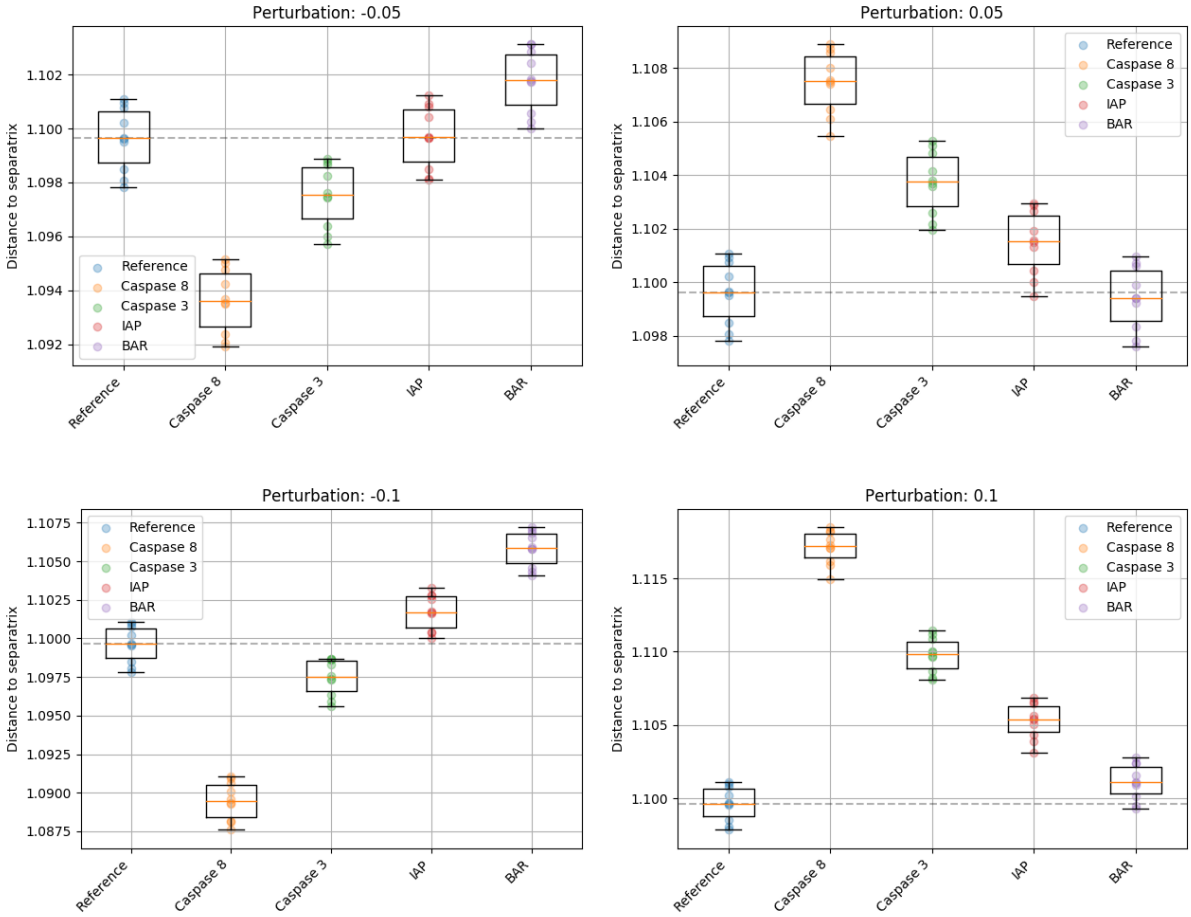
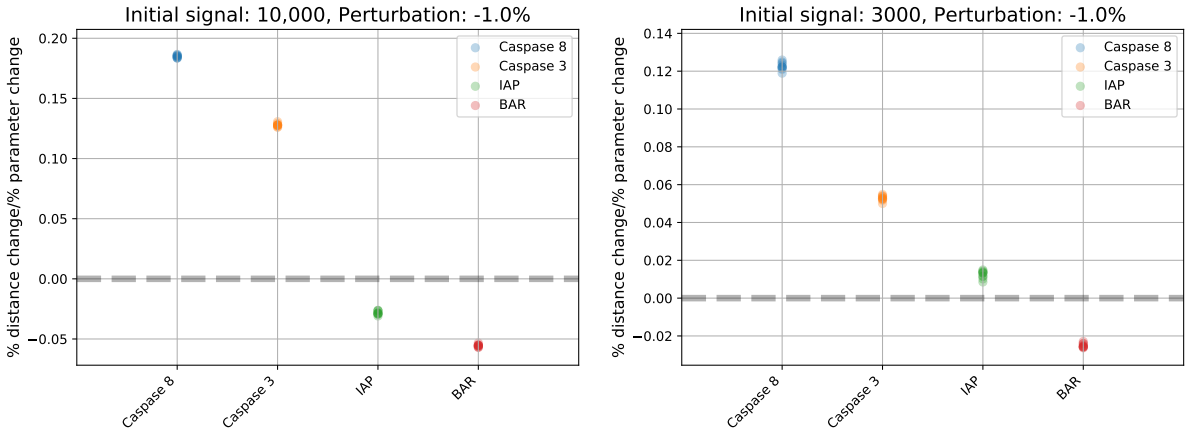


Figure 4.10: Distance from starting point to separatrix surface, calculated as mean distance to all points on the surface. The surface is calculated with 3,000 molecules of initial activated Caspase 8 and each parameter is perturbed: top left; -0.05, top right; 0.05, bottom left; -0.1, bottom right; 0.1 times the initial value. Blue data indicate distances from original starting point to 10 surfaces whereas the rest indicate distances after perturbation of respective production rate parameter.



(a) (b)

Figure 4.11: Percentage mean distance change per percentage parameter change for each separatrix surface of the smaller apoptosis model. 10,000 molecules (a) 3,000 molecules (b) was used as initial activation signal of Caspase 8 and each parameter was perturbed -1% one at a time. From left to right the parameters are production rates of Caspase 8, Caspase 3, IAP and BAR. Using 10,000 molecules as initial activation signal, there is a clear difference in distance between perturbations of the studied parameters and each perturbation is clearly shifting the distance in the direction which would be expected given the function of the variable in the network. Using 3,000 molecules as initial activation signal, perturbation of IAP has an effect opposite to what would be expected.

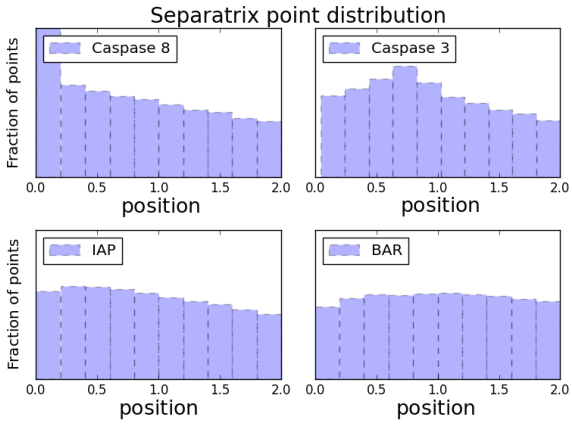


Figure 4.12: Distribution of separatrix points along one axis for surface using 3,000 molecules as initial activated Caspase 8. On the y-axis is the portion of points in each bin of the axis under investigation along the x-axis. Both Caspase 3 and Caspase 8 have more points below the normal value of one, indicating that over a large part of the surface these two nodes are down-regulated, as would be expected. BAR has a more even distribution, indicating that there are many possibilities of crossing the surface where BAR is either up- or down-regulated. Contrary to what would be expected, IAP has more points below one, indicating that even though it is an inhibitor, there is still a larger part of the surface where it is being down-regulated than up-regulated.

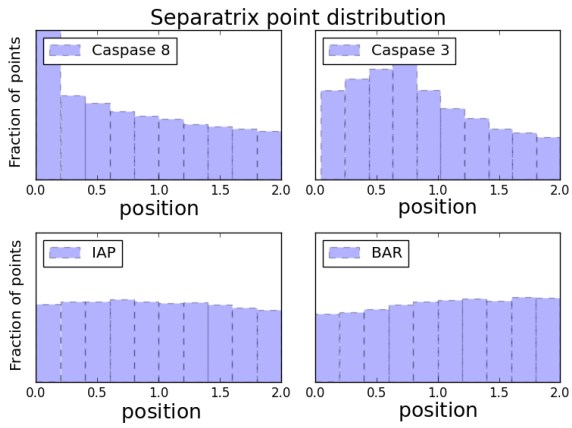


Figure 4.13: Distribution of separatrix points along one axis for surface using 10,000 molecules as initial activated Caspase 8. On the y-axis is the portion of points in each bin of the axis under investigation along the x-axis. The distribution of points on the axis of Caspase 3 and Caspase 8 are even more skewed towards them being down-regulating, indicating that their down-regulation is very important for the chance of crossing over the surface. Compared to the surface for 3,000 molecules of activated Caspase 8, the distribution of both IAP and BAR are skewed towards them having larger values (Figure 4.12).

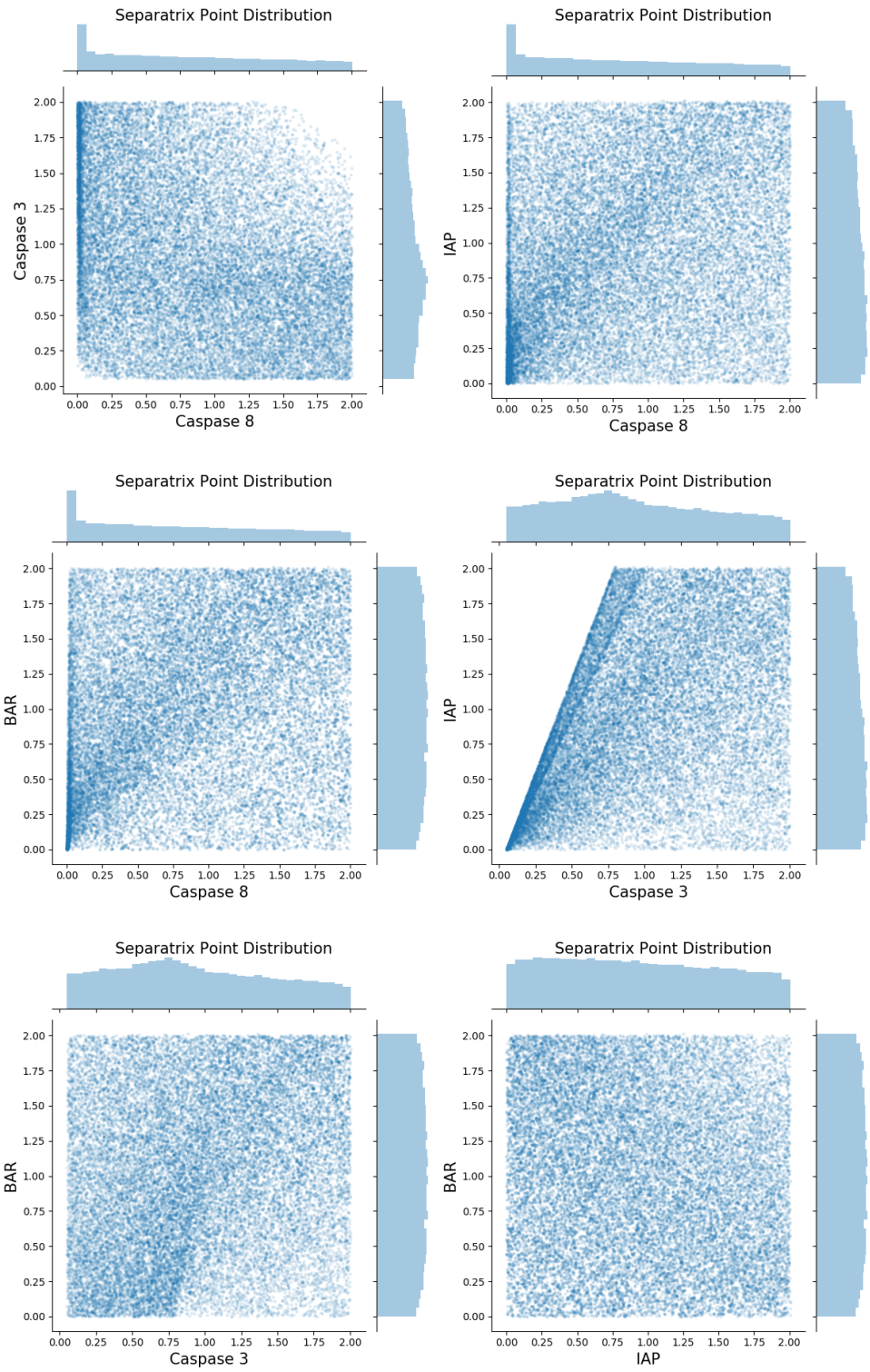


Figure 4.14: Distribution of separatrix points along two axes for surface using 3,000 molecules as initial activated Caspase 8. In almost every case there are points spread over the entire domain.

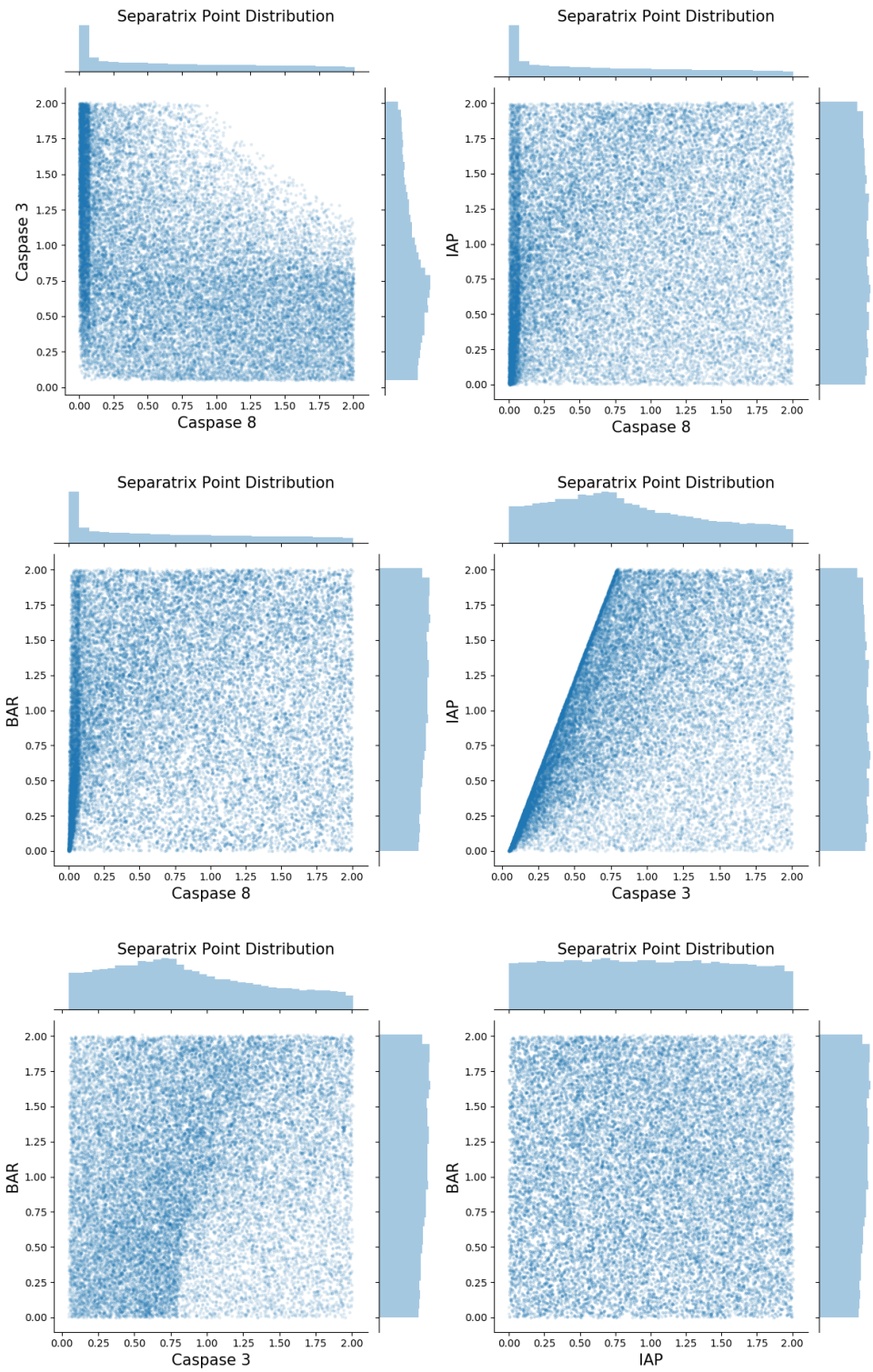


Figure 4.15: Distribution of separatrix points along two axes for surface using 10,000 molecules as initial activated Caspase 8. In almost every case there are points spread over the entire domain. The concentration of points in certain areas is much strong than when using 3,000 molecules as initial activation signal (Figure 4.14)

4.3.2.2 The quality of results for the larger apoptosis model depend strongly on the number of nodes included in the analysis

The larger apoptosis model was run with either all variables whose concentrations are not governed by production parameters, or with just a small subset of those variables around the core of the model, representing the same functional module as the smaller apoptosis model. For the larger set of variables, there was relatively little sensitivity of the variables, with small shifts in the distribution of the distances upon perturbations (Figure 4.16 and 4.17). This lack of sensitivity for some parameters carried over to the distance change of each surface, where even a 0.05 times perturbation in the negative direction resulted in many variables having a distribution of mean distances changes including zero (Figure 4.18c). Decreasing the perturbation to -0.01 times the initial concentration resulted in the distributions of some of the variables to no longer include zero (Figure 4.18a). However, it moved the distributions of some of the more insensitive variables down to zero. A 0.05 times perturbation in the positive direction did result in non-zero overlapping distributions of mean distance changes, but there was very little variation for most of the variables (Figure 4.18d).

Upon inspection of the distribution of the separatrix points along each axis, it was seen that most variables had a close to uniform distribution (Figure 4.19). Two variables with a slightly skewed distribution were FADD and Caspase 8. These variables were also two of the most sensitive ones, as identified in Figure 4.16. Caspase 3 also had a skewed distribution; however, the distance to separatrix did not prove to be any more sensitive to perturbations in this variable than any other of the variables investigated.

When limiting the surface to encompass the variables governing the core of the model around Caspase 3, there was a clearer distinction between the variables. Small negative perturbations of all inhibitors brought the system further away from the surface, whereas negative perturbations of all activators brought the system closer to the surface, as would be expected (Figure 4.20 and G.3). Interestingly the relative order in which the variables affected the distance to the surface corresponded with the order in which the parameters in the smaller apoptosis model affected the distance to that surface, indicating that, although the wiring of the two models are slightly different, they both capture the same type of dynamics. Closer inspection of the distribution of the surface points along each axis showed that a larger part of the surface was located in areas where activators were down-regulated and/or inhibitors were up-regulated than the other way around (Figure 4.21).

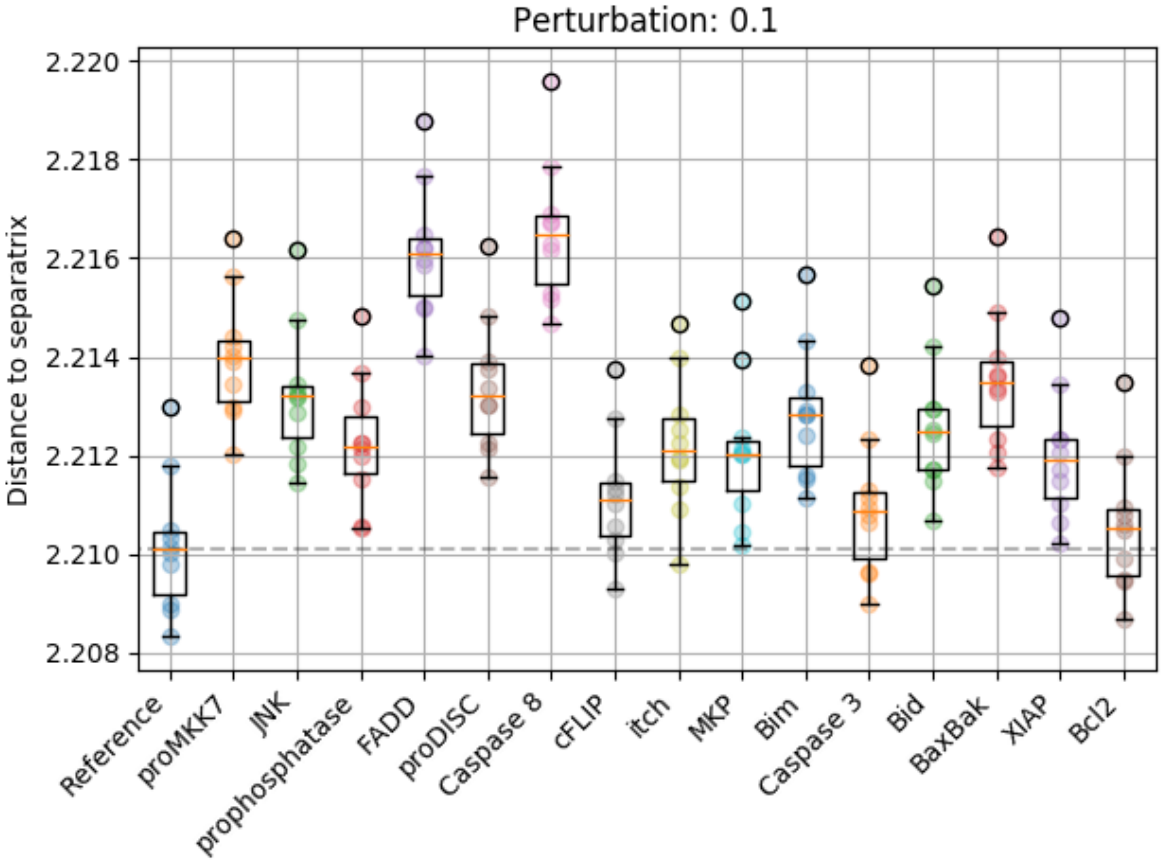


Figure 4.16: Distance from starting point to separatrix surface for larger apoptosis model, calculated as mean distance to all points on the surface before and after each variable is perturbed 0.1 times the initial value. Blue histogram indicate distances from original starting point to 10 surfaces whereas red histogram indicate distances after perturbation.

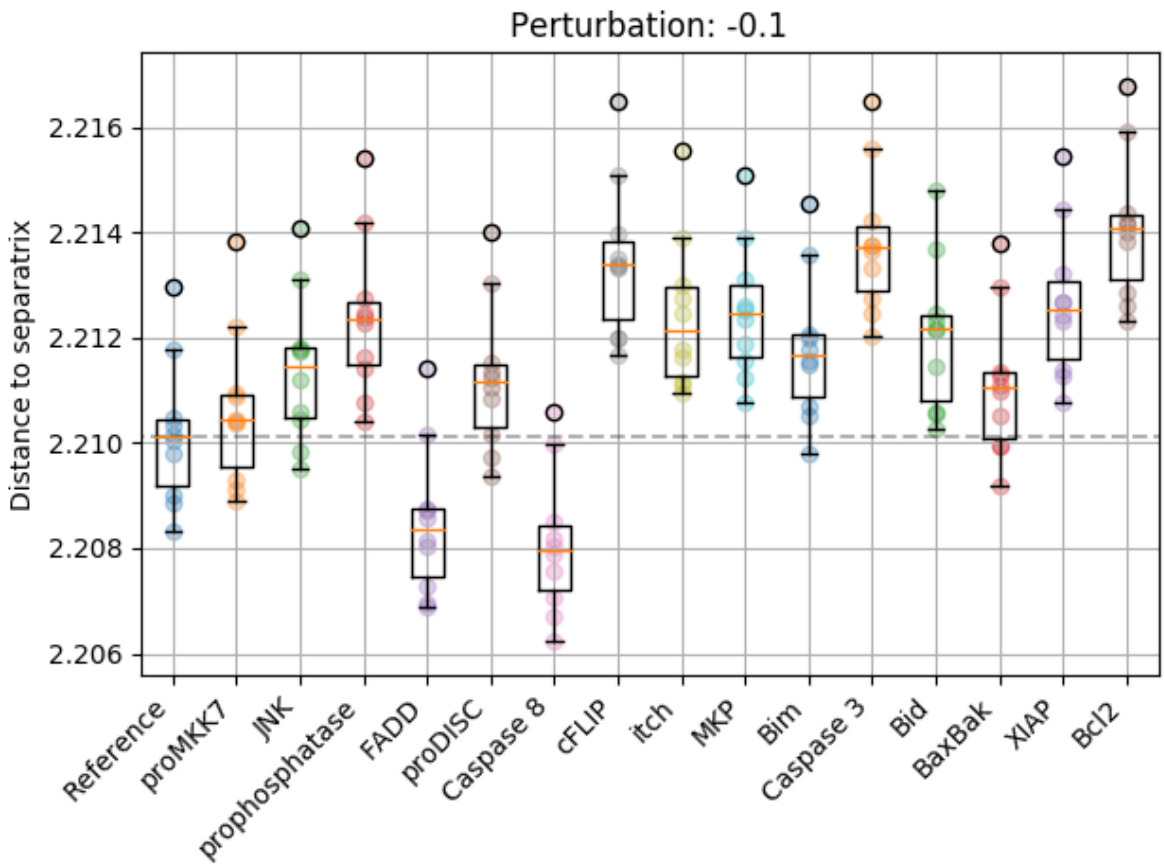
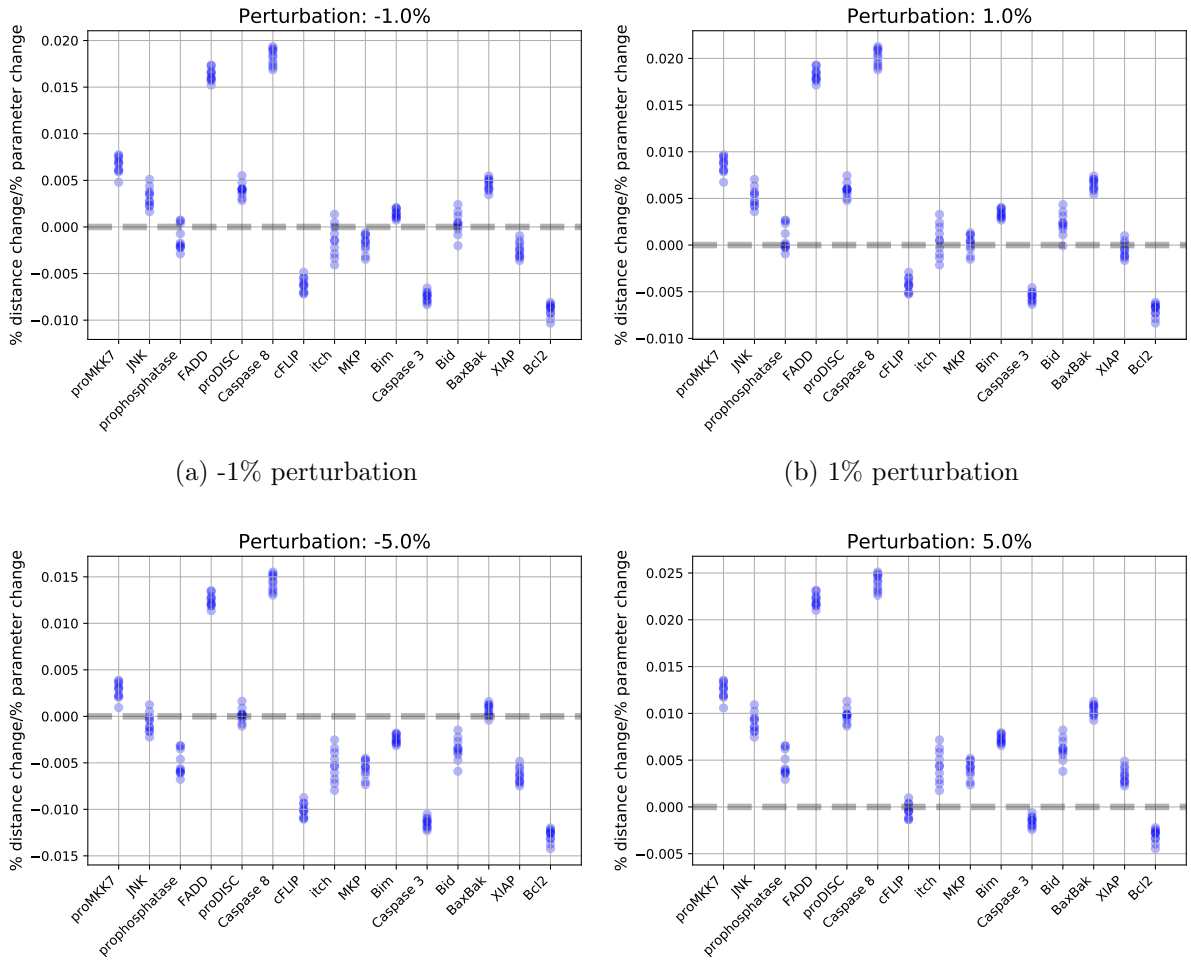


Figure 4.17: Distance from starting point to separatrix surface for larger apoptosis model, calculated as mean distance to all points on the surface before and after each variable is perturbed -0.1 times the initial value. Blue histogram indicate distances from original starting point to 10 surfaces whereas red histogram indicate distances after perturbation.



(a) -1% perturbation

(b) 1% perturbation

(c) -5% perturbation

(d) 1% perturbation

Figure 4.18: Percentage mean distance change per percentage parameter change for each separatrix surface of the larger apoptosis model as each variable is perturbed one at a time, 1% (a–b) or 5% (c–d) times the initial value. (from left to right: proMKK7, JNK, prophosphatase, FADD, proDISC, Caspase 8, cFLIP, itch, MKP, Bim, Caspase 3, Bid, BaxBak, XIAP and Bcl2).

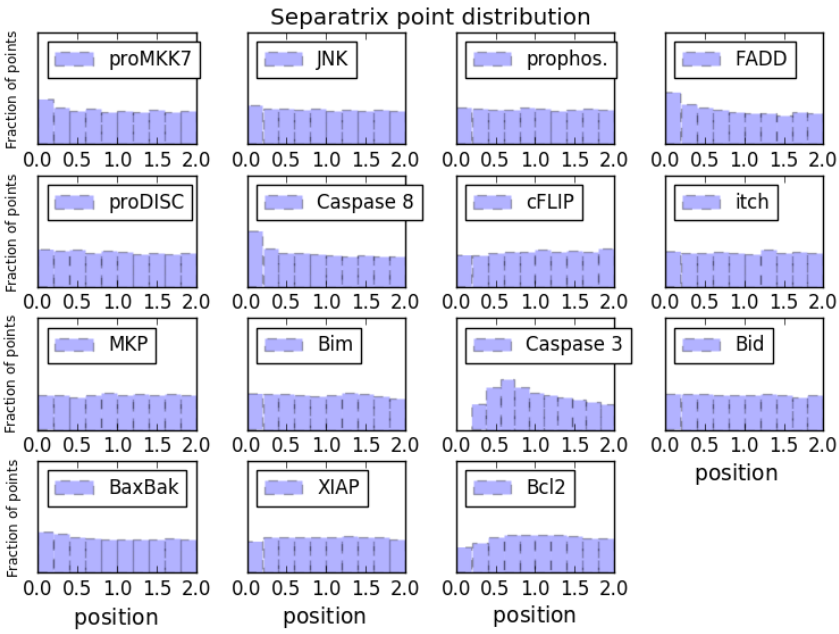


Figure 4.19: Distribution of separatrix points along one axis. On the y-axis is the portion of points in each bin of the axis under investigation along the x-axis. Apart from FADD, Caspase 8 and Caspase 3, the distribution of points are fairly uniform, indicating that they are not more sensitive in on direction than another.

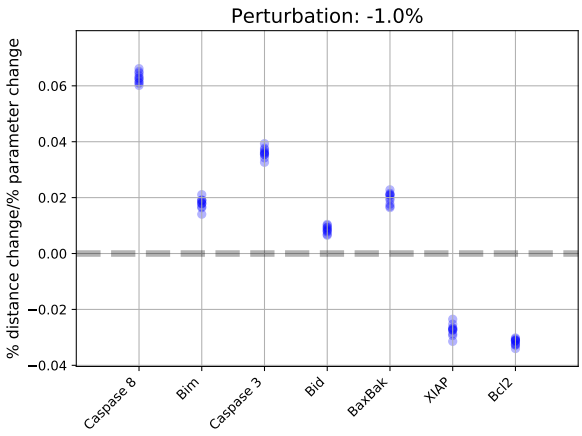


Figure 4.20: Percentage mean distance change per percentage parameter change for each separatrix surface of the larger apoptosis model as each parameter is perturbed one at a time, -0.01 times the initial value. Initial variable values of genes in the core of the model around the Caspase was included in the surface (from left to right: Caspase 8, Bim, Caspase 3, Bid, BaxBak, XIAP and Bcl2). As expected, decreasing any of the activators (first 5 genes in order: Caspase 8, Bim, Caspase 3, Bid and BaxBak) resulted in a shortening of the distance to the surface, whereas a decrease of any of the inhibitors (last 2 genes: XIAP and Bcl2) resulted in an increase in the distance.

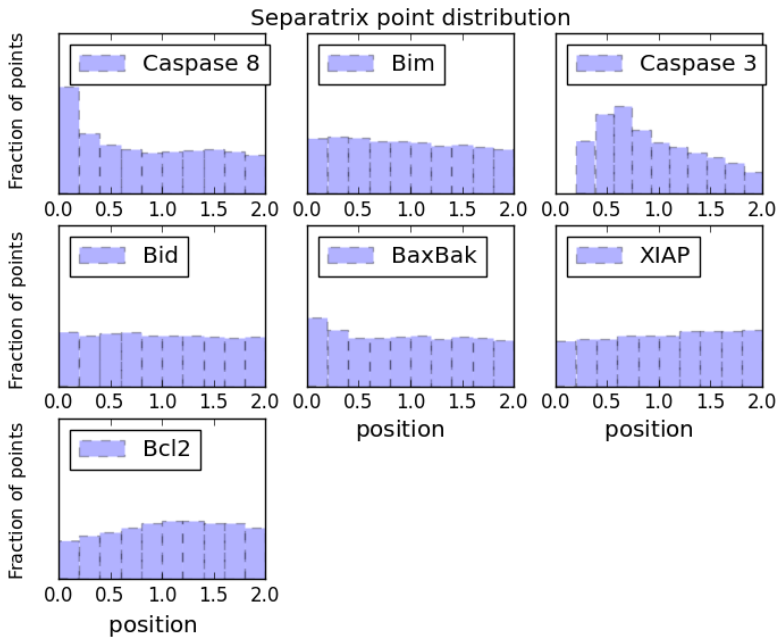


Figure 4.21: Distribution of separatrix points along one axis. On the y-axis is the portion of points in each bin of the axis under investigation along the x-axis. The distribution of points on the axis of activator Caspase 8, Caspase 3 and BaxBak are clearly skewed towards them being down-regulated, whereas the distribution of points for the 2 inhibitors XIAP and Bcl2 are slightly skewed towards them being up-regulated. The distribution of points for the two activators Bim and Bid do not show much of a skew in either direction.

4.3.3 GWAS simulation

4.3.3.1 Risk score correlates strongly with both time to apoptosis and distance to separatrix of the smaller model

Generating only SNPs affecting the rate parameters of the proteins in the model, populations of one million diseased and one million non-diseased individuals were simulated and odds-ratios were calculated according to Section 4.2.2.2. In this way, 50 SNPs with corresponding odds-ratios and a p-value below $5*10^{-4}$ were generated (Appendix Table F.1). Using these SNPs sample populations were simulated and the risk score ($\sum(|\log_e(\text{Odds ratio})| \times \text{number of risk alleles})$ for all SNPs) was correlated to the time to apoptosis as calculated by the model, as well as the distance to the separatrix as defined previously.

When the developed method was later applied experimental data, the smallest population size 50 individuals. As to not overestimate the power of the method the simulated sample size was also limited to 50 individuals. Given the size of the models and the low likelihood of having several SNPs affecting the same gene the number of SNPs used was also limited to 10 SNPs. This sample size was also in agreement with those later used for experimental data. With this setup clear correlations could be identified, both between risk score ratio (RSR, individual risk score/largest possible risk score given the SNPs carried by the individual) and time to apoptosis, as well as distance to separatrix (Figure. 4.22). Repeating the experiment with different sets of SNPs resulted in clear correlations between the RSR and the time to apoptosis (data not shown). The correlation between the RSR and the distance to separatrix also proved robust, although less robust than the correlation between RSR and time to apoptosis (data not shown).

When increasing the number of SNPs in the risk score, without linking them to the model, so that only 10 out of 20 SNPs actually affected the time to apoptosis or the distance, whereas all are used to calculate the risk score, clear correlations could still be seen, although not as strong as when all SNPs had been linked to the model (Figure. 4.23). Repeating the experiment with different sets of SNPs again indicated robust correlation between the RSR and both the time to apoptosis and distance to separatrix, although less robust than when all SNPs had been linked to the model (data not shown).

By varying the number of SNPs in the analysis and the portion of those SNPs linked to the model, a clear correlation could be seen between the fraction of risk score linked to the model (SNPs which affect a parameter in the model) and the significance of the correlation. Moreover, very small fractions of the risk score needed to be attributed to SNPs linked to the model in order to get a correlation with a p-value below 0.01 (Figure 4.24).

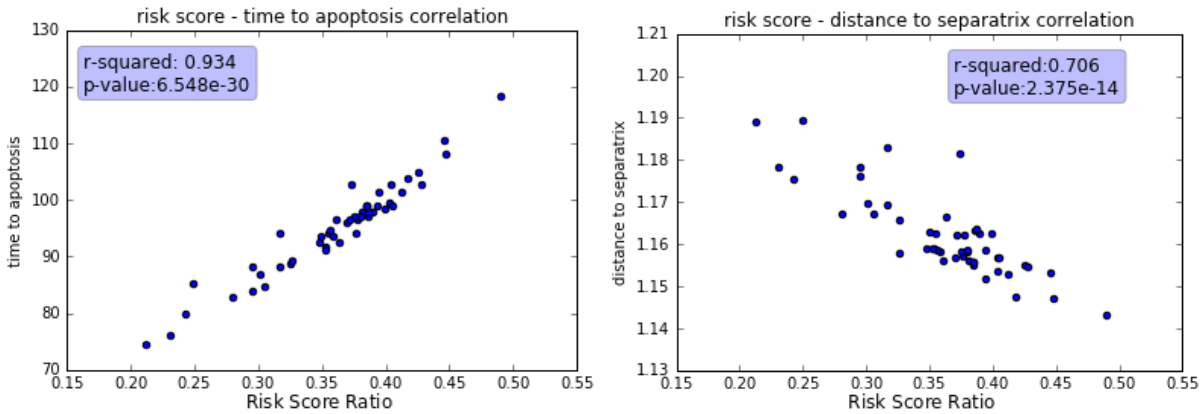


Figure 4.22: Simulations of 50 individuals with generated genotypes for 10 simulated cancer associated SNPs. Left: correlation between RSR and time to apoptosis calculated by applying the SNP perturbations to the corresponding parameters in the model before running it for 5,000 minutes, with 10,000 molecules of initial activated Caspase 8. Right: correlation between RSR and distance to separatrix surface after applying the perturbations to corresponding parameters.

Indeed by repeating the analysis 10,000 times (using 50 individuals each time) and calculating a mean p-value within a moving window of 500 analyses, it could be seen that if around 40% or more of the risk score was linked to the model a significant correlation (p-value < 0.01) was obtained almost every time (Figure 4.25). If the number of individuals was increased to 100 only around 30% of the risk score would have to be linked to the model for similar results. When using the distance to the separatrix instead of the time to apoptosis a similar trend could be seen. However, a much larger fraction of the risk score had to be linked to the model in order to see a significant correlation almost every time (Figure 4.26a). The same was true when increasing the amount of samples in each simulation to 100 individuals (Figure 4.26b).

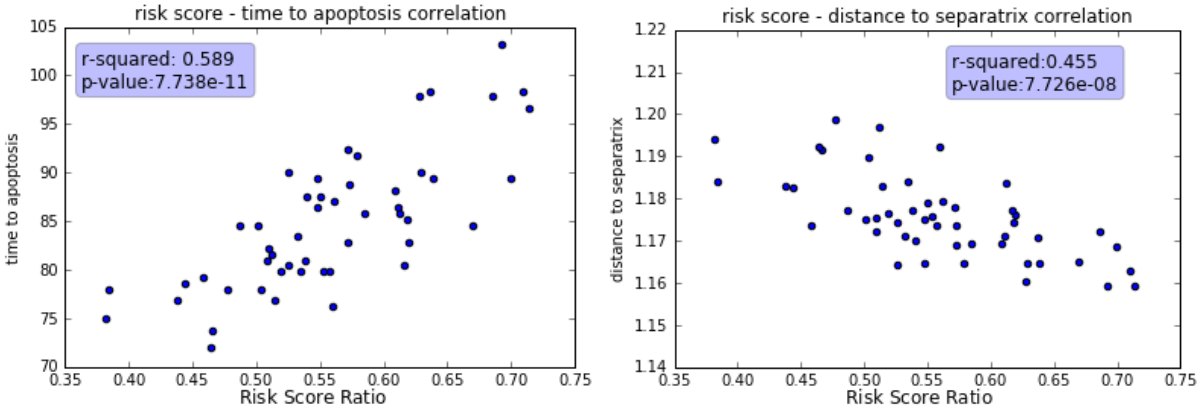


Figure 4.23: Simulations of 50 individuals with generated genotypes for 20 simulated cancer associated SNPs, where 10 were randomly chosen to be associated with the correct parameter, whereas the other did not have any effect on the model. Just as in the previous case in Figure 4.22, there is a clear correlation between the RSR and: left; the time to apoptosis as well as, right; distance to separatrix.

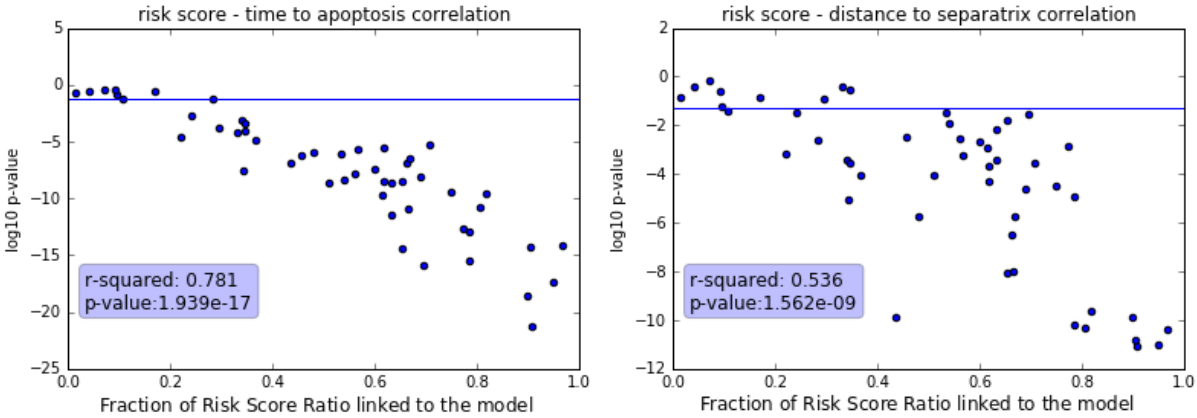
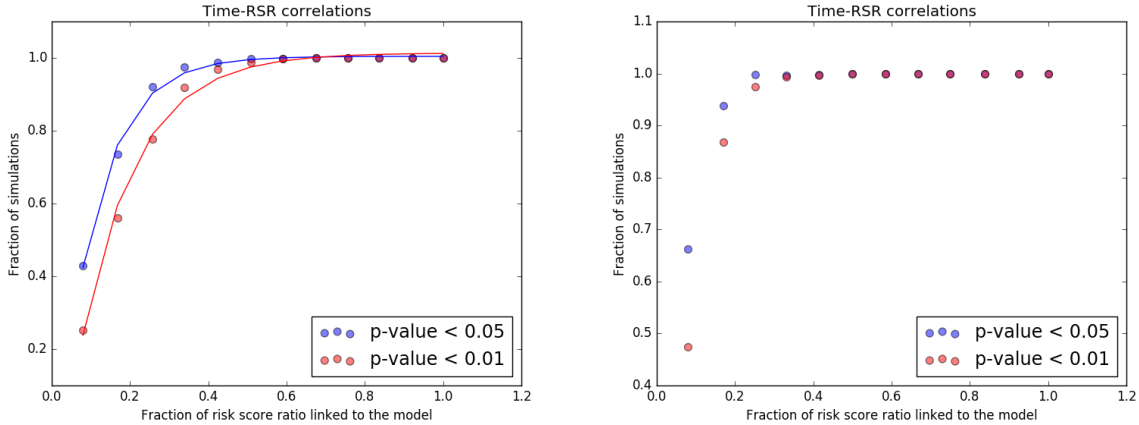
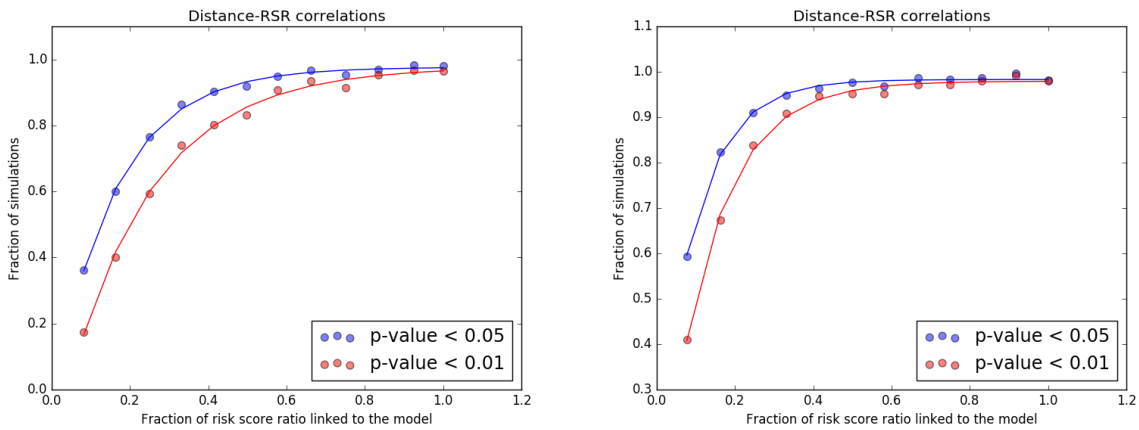


Figure 4.24: 50 calculations of p-values for correlations between RSR and left: time to apoptosis, and right: distance to separatrix. For each experiment 50 individuals were simulated with a random number of SNPs and a random number of those SNPs linked to the model.



(a) 50 samples (b) 100 samples

Figure 4.25: Correlation between time to apoptosis and RSR for the smaller apoptosis model. Ten thousand experiments with either 50 (a) or 100 samples (b) each were generated. In each simulation a random number of SNPs (max 50) were chosen and a random subset of those SNPs were linked to the model. Each point represent the fraction of 500 experiments around that point with a p-value below 0.05 (blue) and 0.01 (red).



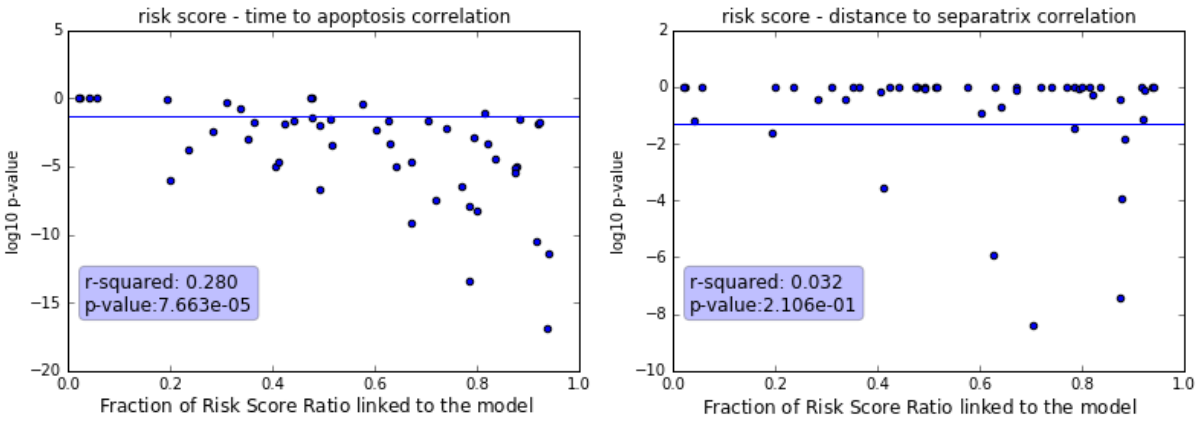
(a) 50 samples (b) 100 samples

Figure 4.26: Correlation between distance to separatrix and RSR for the smaller apoptosis model. Ten thousand experiments with either 50 (a) or 100 (b) samples each were generated. In each simulation a random number of SNPs (max 50) were chosen and a random subset of those SNPs were linked to the model. Each point represent the fraction of 500 experiments around that point with a p-value below 0.05 (blue) and 0.01 (red).

4.3.3.2 The correlation between risk score and the distance to separatrix of the larger apoptosis model depends largely on the number of nodes included

Generating SNPs for all 13 variables used in the larger apoptosis model resulted in 50 SNPs with a p-value > 0.05 (Appendix Table F.2). Whereas the SNPs for the smaller model were spread evenly across all parameters of interest, there was a more skewed distribution of SNPs for the larger apoptosis model. There were no SNPs targeting NFκB or itch and only one SNP targeting proMKK7 or FADD. Also BaxBak as well as JNK and prophosphatase were under-represented with only 2 and 3 SNPs each respectively. The SNPs of these under-represented targets also tended to have a larger effect than many other SNPs. When considering only small effect SNPs (between 0.95 and 1.05 time the initial value), there were only SNPs targeting Bid, Bim, Caspase 3, Caspase 8 and XIAP.

Randomly choosing 10 of the total set of SNPs and linking a random number of them to the larger apoptosis model resulted in a weak correlation between RSR and the time to apoptosis, and an even weaker correlation with distance to separatrix. Repeated simulations with 50 individuals showed a trend of RSR-time correlations with increased significance when the fraction of the RSR linked to the model increased Figure 4.27a). The same trend could not be seen with regards to correlations between RSR and distance to separatrix (Figure 4.27b).



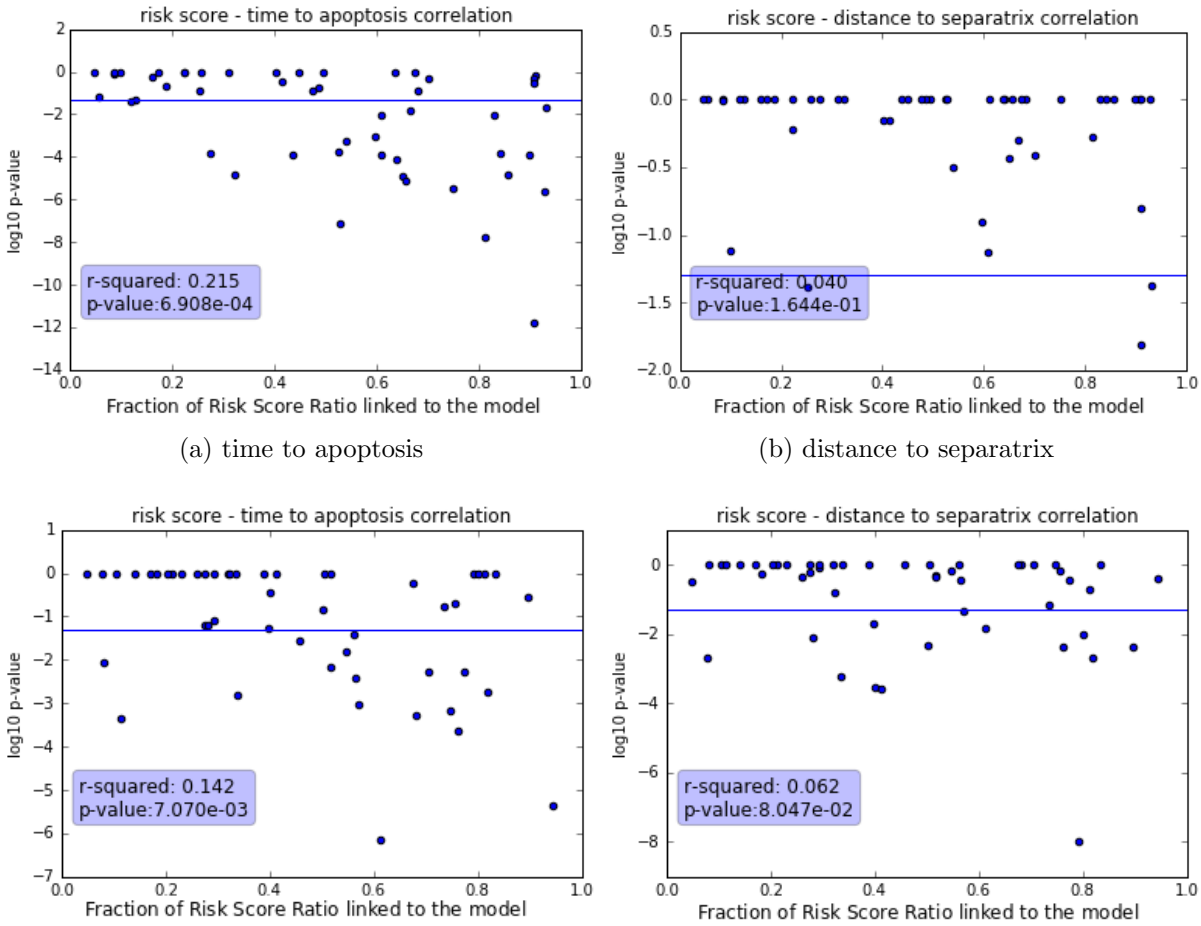
(a) time to apoptosis

(b) distance to separatrix

Figure 4.27: 50 calculations of correlations between RSR and a: time to apoptosis and, b: distance to separatrix for the larger apoptosis model. Each simulations contained 50 individuals with 10 SNPs randomly chosen from the entire set and a random number of those SNPs linked to the model. Both maximum time to apoptosis and separatrix surface was set to 25 hours and all analysed variables were used.

Limiting the SNPs to the ones having a small effect on the variable (between 0.95

and 1.05 time the initial value) left a set of 11 SNPs. When these SNPs were used, the link between increased fraction RSR linked to the model and significance of correlations between RSR and time to apoptosis as well as distance to separatrix became weaker (Figure 4.28a-b). This trend became even more pronounced when limiting the SNPs to effect sizes between 0.97–1.03 times the initial values (Figure 4.28c-d).



(a) time to apoptosis

(b) distance to separatrix

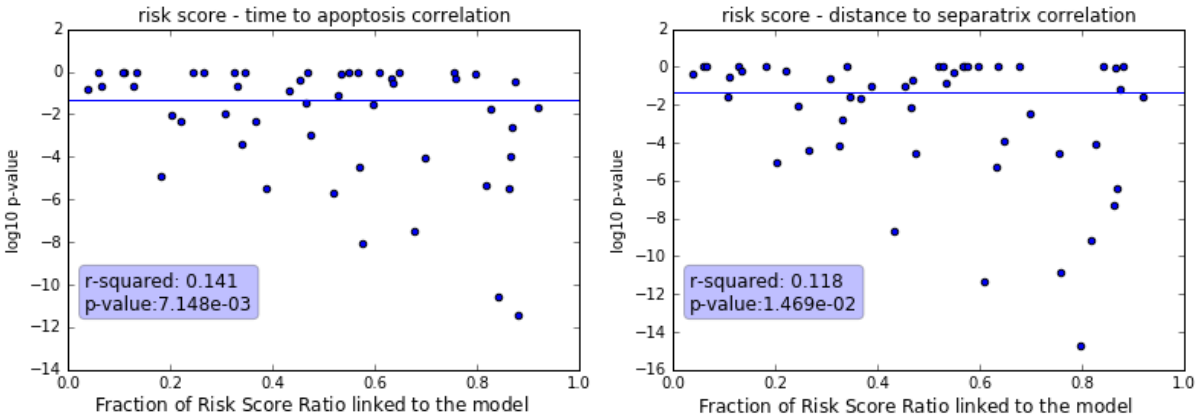
(c) time to apoptosis

(d) distance to separatrix

Figure 4.28: 50 calculations of correlations between RSR and a and c: time to apoptosis and, b and d: distance to separatrix for the larger apoptosis model. Each simulations contained 50 individuals with 10 SNPs randomly chosen from SNPs with a variable effect size between 0.95–1.05 (a–b) or 0.97–1.03 (c–d). A random number of those SNPs were linked to the model. Both maximum time to apoptosis and separatrix surface was set to 25 hours and all analysed variables were used.

The analysis was also performed on a subset of the larger apoptosis model, which included the core proteins BaxBak, Bcl2, Bid, Bim, Caspase 3, Caspase 8 and XIAP, and the separatrix surface was recalculated in the lower dimensional space made up of these

proteins. When considering this smaller separatrix surface, while the trend of correlations between the RSR and time to apoptosis behaved in the same way, the significance of the correlations between RSR and distance to separatrix increased as the effect size of the SNPs was decreased (Figure 4.29 and 4.30).

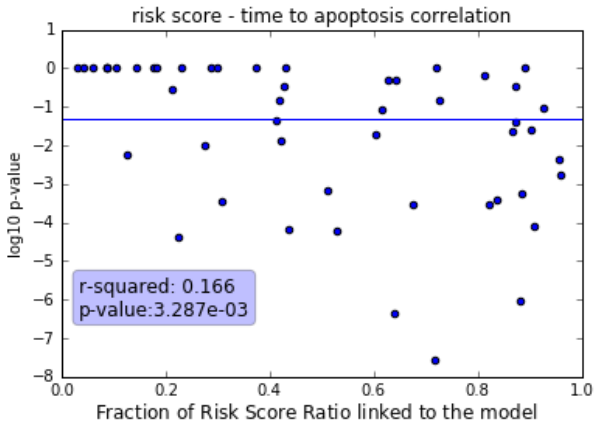


(a) time to apoptosis

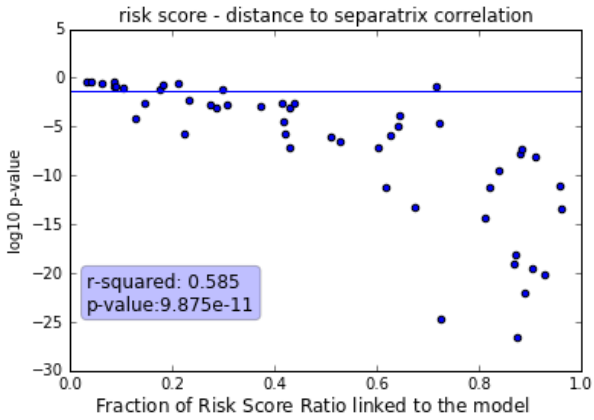
(b) distance to separatrix

Figure 4.29: 50 calculations of correlations between RSR and a: time to apoptosis and, b: distance to separatrix for the larger apoptosis model. Each simulations contained 50 individuals with 10 SNPs randomly chosen from the entire set. Out of these SNPs, a random number were linked to the model. Both maximum time to apoptosis and separatrix surface was set to 25 hours and only variables around the Caspase signalling were used.

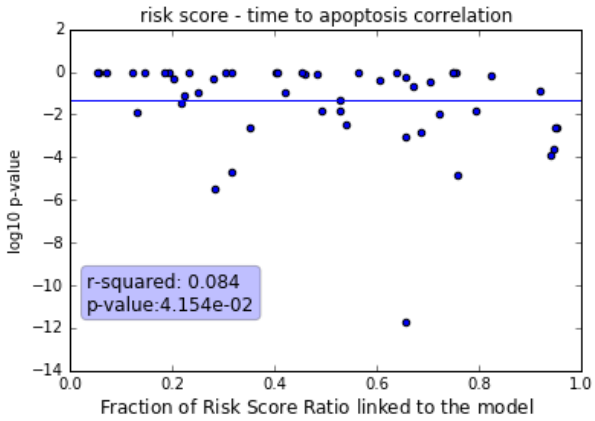
To test whether the difference in trend patterns was due to the surface being to close to the starting point and larger mutations bringing the system too close, the same analysis was repeated using the smaller variable set (excluding BaxBak) for three time lines; 25, 50 and 100 hours, either sampling from the entire set of SNPs or limiting to those with an effect size between 0.97–1.03. However, there was no clear change in behaviour as the time for the separatrix surface was increased (Appendix Figure G.6 and G.7).



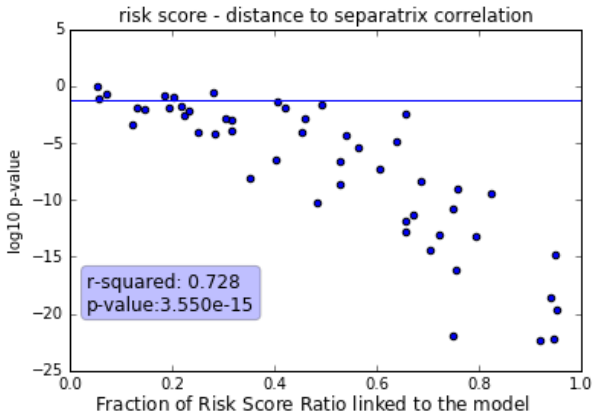
(a) time to apoptosis



(b) distance to separatrix



(c) time to apoptosis



(d) distance to separatrix

Figure 4.30: 50 calculations of correlations between RSR and a, and c: time to apoptosis and, b, and d: distance to separatrix for the larger apoptosis model. Each simulation contained 50 individuals with 10 SNPs randomly chosen from SNPs with a variable effect size between 0.95–1.05 (a and b) or 0.97–1.03 (c and d). Furthermore, a random number of those SNPs were linked to the model. Both maximum time to apoptosis and separatrix surface was set to 25 hours and only variables around the Caspase signalling were used.

4.4 Discussion

In this chapter the concept of relating the potential for phenotype change of a system to the difference between the starting configuration of the system and a set of potential configurations causing that phenotype change was introduced. These potential configurations were represented by a separatrix surface in parameter and/or variable space, separating configurations of the system which exhibit the two different phenotypes. It was hypothesised that the mean distance between the starting position in the space and the points on this surface could be used as a proxy for the risk of developing cancer. A method of generating these separatrix surfaces was also presented and the method was applied to a larger and a smaller apoptosis model, using both single perturbations with known effect on the systems and simulated population studies, relating the distance to separatrix and the time to apoptosis to the risk score of simulated genotypes, affecting large parts of the model at the same time.

By altering the number of initial points in parameter space and the number of cycles for identifying points on the separatrix surface it was shown that it is relatively straightforward to decide whether the level of detail of the generated surface is sufficient. Moreover, it was shown that the precision of the surface was much more dependent on the number of initial points than on the number of cycles used for identifying the surface (Figure 4.3 and 4.4). By generating a couple of surfaces with an increasing number of points, it would be very easy to decide on a final surface by considering the rate at which the standard deviation of the distance to the surfaces declines. For this analysis, all surfaces were generated at the same time and for best results, the largest surface was used for further analysis. However, for any future models, a criteria could be formalised so that points would be added until the rate of decline in standard deviation reached a certain level. This would not only make the decision of generating a surface less arbitrary, but also possibly make the results between models more comparable.

Using small perturbations of single parameters within the smaller apoptosis model, the method was shown to be adept at relating differences in parameter space to shifts in distance to the separatrix surface. Although there was a great overlap of distances between the normal and perturbed states, the difference became more clear when comparing the difference in distance within each surface separately.

The individual perturbations also revealed limitations in the method. One of these was the importance of defining the separatrix beforehand in a way that it was sufficiently far away from the starting point. When the activation signal was set to 3,000 molecules instead of 10,000, rendering the whole system more sensitive to further perturbations, a

small perturbation in IAP did not have the expected effect (Figure 4.11). It did, however, behave as expected when the initial activation was increased to 10,000 molecules. For IAP, both when the initial activation was set to 3,000 and when it was set to 10,000, the effect on the distance upon larger perturbations was also contrary to what would be expected given the function of the protein in the model (Figure 4.8 and 4.10). However, the effect was much more prominent with the lower activation signal. For the lower activation, large positive perturbations also resulted in a shift in distance contrary to what would be expected of an inhibitor, again highlighting the importance of designing the experiment so that the separatrix surface is far away from the initial starting site in parameter space.

This difference in behaviour of IAP and BAR could be explained by looking at the histogram of separatrix surface points projected onto each axis (Figure 4.12 and 4.13). Both Caspase 3 and Caspase 8 had a point distribution skewed below 1.0, meaning that a positive perturbation would bring the system further away from more points of the surface than the number of points it would bring the system closer to. Conversely, a perturbation in the negative direction would bring the system closer to a larger part of the surface. This is what can be seen using both amounts of initial activation signal, with an increased distance for positive perturbations and decreased distance for negative perturbations. However, for the surface with the lower activation signal, the points along the IAP axis are skewed towards smaller values (Figure 4.12), contrary to what would be expected of an inhibitor. This means that even though the system takes longer to respond when IAP in increased, it brings the system further away from a large amount of possible combinations of perturbations which would cause the system to shift phenotype. A small negative perturbation also brings the system closer to a large amount of points. However, as the perturbation increases, the amount of points which it is moving further away from also increases, resulting in a limited effect. When using the the larger activation signal, the distribution of points is more even and the effect is not as big. Likewise, the distribution of points along the BAR axis is slightly skewed downwards when using the lower activation signal, but skewed upwards when using the higher activation signal. Consequently, the unexpected behaviour can only be observed when using the lower activation signal.

The reason for the points not being evenly distributed along all axes is because of the curvature of the surface. For example, when Caspase 8 is low, the effect that has on the system can not be compensated easily by other perturbations in any other gene. Consequently, there will be many combinations of perturbations of all genes which will give almost the same outcome of the model. When all of these points are projected down on the Caspase 8 axis, this manifests as a large peak at the lower end of the distribution.

When the same analysis was performed on the larger model, the results were naturally more complicated. In general the system was less sensitive to single perturbations, shown both by the distributions of distances upon perturbations (Figure 4.16 and 4.17) and the change in distances (Figure 4.18). Also the distribution of points along each axis of the variable space was more evenly distributed compared to the smaller apoptosis model (Figure 4.19). This could be due to the scale of the model and that any perturbation could be compensated by a combination of perturbations spread over a larger amount of genes, compared to the smaller model.

When the separatrix surface of the larger model was confined to the lower dimensional space constituting to genes around the core of the apoptosis network, more resembling the smaller apoptosis model, the system became more sensitive to perturbations (Figure 4.20 compared to 4.18, keep in mind that the magnitude of perturbation is different in the two graphs). More importantly though, in this lower dimension space, perturbations in all single nodes resulted in shift in distance corresponding to what would be expected given the function of the protein in the network. When using the larger set of nodes, negative perturbations in most activators (resulting in slower response time) had resulted in an increase in distance (Figure 4.18). Likewise, positive perturbations in inhibitors had either resulted in an increase of distance (XIAP), or a distance change distribution overlapping zero (Bcl2)).

From the distributions of points on both types of surfaces it appears that the nodes with a skewed distribution in either direction, mostly correspond to the nodes also present in the subset of the model. It is likely that this, unintentional, filtering out of less sensitive nodes resulted in a more sensitive system overall. It is also possible that the heightened sensitivity is due to the fact that as the number of dimensions increase, the contribution of each dimension to the length of a vector in that space, decreases (The length of a vector is defined as $\sqrt{\sum_{i=1}^{n} x_i^2}$. As n increases the contribution of x_i on the length decreases.).

This difference in sensitivity between nodes of the larger apoptosis model could also be seen when examining the simulated phenotype shift associated SNPs generated. There was a clear over-representation of SNPs targeting the subset of nodes and two outer nodes in the larger model (NFκB, and itch) not having a single SNP targeting them.

This general lack of sensitivity might explain the weak correlations between the RSR in the simulated populations and the distance to separatrix in the larger implementation of the larger apoptosis model (Figure 4.28) and why a correlation was more visible using the subset of the model (Figure 4.30). Interestingly, these trends were only visible when using only the set of SNPs with small effects of the genes in the model. The effects were in fact so small that the RSR could not be reliably correlated with the time to apoptosis as

calculated by the model. Conversely, when the effects were large enough to be correlating the with the time to apoptosis, they did not correlate with the distance to separatrix. This is probably due to the narrow span of the variable space (0 to 2 times initial value) and the larger perturbations probably pushed the system so close to the boundary that the surface was no longer representative of the potential mutations of the system.

Interestingly, the same trend could not be seen when limiting the SNPs of the smaller apoptosis model to the ones with the smallest effect on the parameters. When using all SNPs there was a clear correlation between RSR and both time to apoptosis and distance to separatrix (Figure 4.24). However, even when only considering SNPs with a perturbation effect $0.97 < x < 1.03$, there was a clear trend of correlations with increased significance, both when correlating RSR to time to apoptosis and distance to separatrix (data not shown).

Furthermore, these correlations between RSR and both time to apoptosis and distance to separatrix could be identified with high accuracy, even when very small fractions of the RSR ratio was actually linked to the model (Figure 4.25 and 4.26). This gives increased hope that the method will be possible to use on real experimental data, when it is not always known what a SNP is doing and there is bound to be more noise on all levels of measurement.

The differences in results between the models and between settings of the same model, point towards the need for a more in-depth investigation into how the size of the model, and the network dynamics, affect how suitable the method is. Further research would also need to be performed on how the limits of the separatrix space affects the extent to which the method represent actual biological development. However, the fact that separatrix was able to correlate with the RSR in settings where the standard output of the model failed, shows the potential of the method and that it has a possible use case which can not be covered by standard sensitivity analysis of system dynamics.

Chapter 5

Linking experimental data to model behaviour and separatrix surface

In chapter 4 we examined the effects of a theoretical mutation on the output of a dynamical model when interpreted as a change in a parameter value. We saw that the resulting change in model behaviour could be linked to the risk of the system to change model behaviour from what could be considered normal to what can be classified as diseased (in the case of the apoptosis model, this was defined as a time to activation of apoptosis which would render the cell effectively non-responsive). This was based on the biological assumption that an individual will accumulate mutations randomly in all genes and eventually this accumulation will cause the cellular network in which the gene operates to break down or change behaviour so much that it will exhibit a phenotype associated with disease. Under this assumption, Single Nucleotide Polymorphisms (SNPs) affecting production rate parameters of the four genes in a smaller apoptosis model by Eissing *et al.* [95] and initial concentrations of the products of 15 genes in the larger apoptosis model by Schlatter *et al.* [91] were examined. Using simulated SNPs with known odds ratios with respect to disease phenotype and a specified effect on the gene target in the model, it could be shown, not only that the risk score of individuals with random genotypes of these SNPs could be correlated with increasingly disease like model behaviour, but that the risk score could also be correlated with the distance from starting position to the separatrix in the parameter or variable space. This meant that the effect the genotype had on the model in non-diseased state could directly be linked to the increased risk of the system to acquire a diseased phenotype later on, following mutation events.

In this chapter the framework developed in chapter 4 will be extended and applied to experimental data from human cell lines and tissues. Since a large part of the cancer associated SNPs are located outside of protein coding regions and it is thought that they affect RNA expression levels[30, 45] and consequently protein levels, RNA sequencing data

will be used as a proxy for protein expression. Due to limited knowledge about what effect known disease-risk associated SNPs which do not affect RNA expression levels have, the analysis will be limited to differences in protein expression. Since the results in chapter 4 were clearer in the smaller apoptosis model by Schlatter *et al.* compared to the larger model, only the smaller model will be used in this study.

RNA expression values will be used as proxies for concentration of the proteins in the model and time to apoptosis as well as distance to separatrix will be related to the risk score (cumulative risk of all SNPs under investigation) of the individual. SNPs associated with breast and prostate cancer will be examined, using both the dataset as a whole and by extracting a subset of strong candidates, likely to affect the model under investigation. The SNPs will be extracted and combined with RNA sequencing data of both lymphoblastoid cell lines from the 1000 genome project and GEUVADIS, and of normal breast and prostate tissue from The Cancer Genome Atlas (TCGA). TCGA is a project with the goal to identify and characterise genetic mutations responsible for development of cancer in various tissues.

Both breast and prostate cancer have been extensively researched and there is a strong hereditary aspect to risk of developing these cancers, with many SNPs known to be associated with this risk. For the same reason, there is also more tissue specific data available for these cancers, compared to many other tissues.

For the lymphoblastoid cell lines, the breast cancer associated SNPs will be used. Since these samples are not breast tissue derived, there is no guarantee that the breast cancer associated SNPs will have the same function in these cell lines as they do in the process of breast cancer development. However, if a correlation can be found it could potentially be validated in vitro. This could be the basis of a very useful model, where results from simulations could inform about and guide towards hypothesis formations which could be tested in the biological system and vice versa.

5.1 Materials and Methods

5.1.1 Data Collection and Preprocessing

From a paper published by Michailidou *et al.* 2017 [45], 177 breast cancer associated SNPs and genes mapping to them were collected (Appendix Table H.1). A further 142 prostate cancer associated SNPs and genes mapping to them were collected from a paper published by Schumacher *et al.* 2018 [32] (Appendix Table H.2). These publications also contained the odds ratios (the strength of association between a genotype and the development of

the specific cancer) which were used to calculate the risk score ($\sum(|\log_e(\text{Odds ratio})| * (\text{number of risk alleles}))$) for each individual during this work.

As a first data set, genotype data was acquired for the set of cell lines denoted GBR and CEU in the 1000 Genome Project [105] [59]. These are lymphoblastoid cell lines (B cells that have been immortalised through Epstein-Barr virus transformation) collected from individuals in Great Britain and a population in Utah with central European ancestry, respectively. Each line of the data contains, among other things, information about the reference allele and alternative allele, genomic position, rs-ID (a unique SNP identifier issued by dbSNP (https://www.ncbi.nlm.nih.gov/snp/)) and genotype for each individual in the study. RNA expression data for the same cell lines was collected from the GEUVADIS project and contained Reads Per Kilobase (of transcript) per Million (of reads) (RPKM) values for each gene [106].

For the breast and prostate SNP analysis, genotype and RNA expression data from TCGA was gathered from the GDC Data Portal [107] (genotype data was collected from the legacy archive [108]). The breast tissue samples were restricted to female origin and for both tissue types only normal tissue was selected. To further minimise the diversity in genotype-phenotype association, the individuals were limited to those characterised as white. This was done to mimic the selection of European individuals or individuals of European descent which was done for the lymphoblastoid cell lines.

In all three cases the RNA expression data consisted of RPKM values for the four genes XIAP (ENSG00000101966.11), BFAR (ENSG00000103429.9), CASP8 (ENSG00000064012.20) and CASP3 (ENSG00000164305.16).

The TCGA genotyping had been performed on Affymetrix Genome-Wide Human SNP Array 6.0 chips and came encoded with the chip-tag IDs for each SNP. Using data from Thermo Scientific's webpage [109, 110] the tags were mapped to dbSNP rs-IDs.

To increase the amount of SNPs being used in the analysis, SNPs in linkage disequilibrium (LD) with data set SNPs, meaning that there is a non-random association between them, were also considered. For all 177 breast and 142 prostate cancer SNPs in the analysis, any SNP within 50,000 base pairs of a SNP in the original data set and in LD ($r^2 > 0.8$) with that SNP was collected from PLINK [111] and filtered to only contain SNPs which were present on the chips used. This resulted in a 418 SNPs linked to the breast cancer associated SNPs and 378 SNPs linked to the prostate cancer associated SNPs.

For both the breast cancer and prostate cancer analysis, for each SNP in the expanded data set a tag (sequence on the chip used to capture the DNA) was identified in the chip data from TCGA. In the case that the original SNP was covered by the chip, this tag was

used. Otherwise the tag corresponding to the SNP with the highest LD with regards to the original SNP was used.

Out of the filtered SNPs only those for which genotype data was available in all samples were kept resulting in 90 SNPs for the breast cancer tissues and 86 SNPs for the prostate cancer tissues. The Risk Score Ratio (RSR) was then calculated by summing up the absolute of the $\log_e$-Odds ratio (risk score) times the number of risk alleles for each individual and dividing it by the maximum possible risk score, that is the theoretical risk score an individual would have if they had two alleles of all risk associated SNPs used in that analysis.

5.2 Model

The small apoptosis model published by Eissing *et al.* (2004) [95], implemented in Section 3.2.1.1 (Appendix Equation B.1-B.8 and Table B.1 and B.2) and used in chapter 3 and 4 was further used in the work of this chapter. The four parameters corresponding to production rates of the four proteins Caspase 3, Caspase 8, IAP and BAR were targeted for further analysis. The initial activated Caspase 8 signal was set to 10,000 molecules and was subtracted from the total amount of inactive Caspase 8 before starting the simulation.

For all analysis in this chapter, the RNA-seq data corresponding to the four genes CASP3, CASP8, XIAP and BFAR was used to adjust the start position in parameter space as described below. When parameters were perturbed from the standard values (the values used in the original paper), the system was first run to 5,000 minutes with initial Caspase 8 activation set to zero. This meant that no apoptosis signal was being transmitted through the network and the system was allowed to find its new steady state, with concentrations in agreement with the new production parameters. The final concentrations reached after this run were then used as initial concentrations in the actual run containing activated Caspase 8 from time 0. The model was run up to 5,000 minutes after initial activation signalling and the time to onset of apoptosis, interpreted as time when Caspase 3 activation signal reached 1,000 molecules, was measured.

5.2.1 Normalisation of RNA-expression values and fitting of parameters to expression values

To "fit" the RNA-seq data to the model in question the mean RNA expression level of all individuals in the analysis for each of the genes was assumed to correspond to the published values of the corresponding parameter. Furthermore, the parameter value is assumed to relate to the RNA expression value as:

$$P_{i,j} = \frac{X_{i,j}}{mean(X_i)} \tag{5.1}$$

$$\text{where: } X_{i,j} = R_{i,j} - 0.95 \times (R_{i,j} - mean(R_i)) \tag{5.2}$$

where:
$P_{i,j}$ is a positive expression level of protein i for individual j
$X_{i,j}$ is transformed RNA expression value of gene $R_{i,j} > 0$

For each individual, 95% of the difference between that expression value and the mean value was subtracted, reducing the spread of the expression levels. The level 95% was chosen under the assumption that all normal cell should be responding to apoptotic signalling and subtracting more than 95% meant that some samples became non-responders. Each value was then divided by the mean to produce a distribution centred around one. Before running the model, each parameter under investigation was multiplied with the perturbations (normalised RNA expression values). This resulted in all of the individuals showing a behaviour which could be considered to fall within a range of normal behaviour, that is, they all responded to apoptosis signalling within the set time frame. The resulting time to apoptosis for each individual was then correlated with the risk score ratio made up of cancer associated SNPs.

5.2.2 Separatrix analysis

The RNA-expression values of the four genes XIAP, BFAR, CASP8 and CASP3 were normalised as described above and corresponding parameter values were perturbed as previously described before applying the separatrix analysis method explained in Chapter 4. In brief, the method defines a criterion for the system output at which it no longer can be classified as a normally behaving system. Any instance of the system falling within this criterion is considered normal and any system which falls outside is considered diseased. In the case of the smaller apoptosis model the criterion for normal behaviour was set to an activation of Caspase 3 of at least 1000 molecules, within 10,000 minutes. The result of this criterion is that a surface forms in parameter space separating systems of normal behaviour from systems of diseased behaviour. The method calculates the average distance from points on this surface to the starting position of each sample after applying the perturbations inferred from the RNA-expression data. The separatrix surfaces used were the same as those used in chapter 4.

5.2.3 Statistical analysis

In this work regression analysis was performed between RSR and RNA expression, time to apoptosis, and distance to separatrix. In addiction regression analysis was performed between RNA expression levels and time to apoptosis. In all cases the p-value was used to measure significance of the correlation. This approach has many known limitations. One problem is that the p-value of a correlation depend both on the magnitude of the association and on the sample size. This means that even if the correlation is very small it is possible to get a significant p-value with a large enough sample size. Conversely, even a strong correlation may not result in a significant p-value if the sample size is very small.

Null hypothesis significance testing also doesn't give support for any alternative hypothesis in the case of the null hypothesis being rejected. If the level of support for either of the two hypotheses is of value, it is better to use alternative methods, such as Bayes factors which assess the support of the data for one hypothesis over another. However, using Bayes-Factors require the formulation of two hypotheses with prior distributions and it was not clear what these priors should be in the case of the work performed in this chapter. Furthermore, since this the work presented here is a novel method and the available data is very limited, leading to likely low statistical power, the main interest was to see if the correlation coefficient between the two variables was zero or not. It was therefore concluded that reporting p-values was sufficient.

5.2.3.1 Correcting for multiple tests

One of the most common ways to correct for family wise error rate is by performing a Bonferroni correction, where the critical p value is adjusted by dividing α with the number of tests performed. However, due to the large number of tests performed in this study combined with the low sample size and the noisy nature of biological data, a Bonferroni correction would most likely be too conservative and result in any true correlations almost certainly being thrown away.

Instead a correction for the False Discovery Rate (FDR) was conducted according to the Benjamini-Hochberg procedure, where a certain FDR is decided to be acceptable and the critical p-value is adjusted, accepting that a given percentage of the significant results will be false positives.

All p-values were ranked from lowest to highest and the Benjamini-Hochberg critical values was calculated:

$$p_i \leq \frac{i}{m} Q, \tag{5.3}$$

where i is the rank, m the total number of familywise tests and Q is the false discovery rate.

As a test family all correlations with RSR and RNA expression, time to apoptosis or distance to separatrix were considered for all 3 tissue types, resulting in 42 tests. Because of the low statistical power, due to the sample size and the noise in the data Q was set at 0.25 to minimise the risk of rejecting true positives.

When applied to the family of tests, this resulted in the threshold of significance being $p \leq 0.059$.

The test between RNA expression and time to apoptosis were considered belonging to a separate test family and corrections were conducted separately and the threshold of significance for the 4 tests ($Q = 0.25$) was calculated to be 0.004.

5.3 Results

5.3.1 Breast cancer associated SNPs do not correlate with model behaviour for lymphoblastoid cell lines

From the 1000 genomes project, 154 individual lymphoblastoid cell lines were identified, for which genotype data and RNA expression data were available from 1000 Genomes Project and GEUVADIS respectively [112, 106]. Out of these cell lines, 76 were of female origin and were chosen for further analysis. After normalisation, the RNA expression values were still skewed with a slight tail extending into higher expressions for XIAP, CASP8 and CASP3 (Appendix Figure I.1). Only BFAR had a normal distribution according to the Shapiro-Wilk test for normality [113] (p-values: XIAP, 4.0×10^{-09}; BFAR, 0.45; CASP8, 1.6×10^{-10}; CASP3, 4.6×10^{-6}). Using this normalised RNA expression data to adjust the protein production parameters in the model before simulation, the distance to the separatrix and time to apoptosis as calculated by the model was compared to the RSR of each sample and a regression analysis was performed.

No significant correlation was identified between the RSR and the time to apoptosis (Figure 5.1, p-value: 0.67, r^2: 0.002) or between the RSR and the distance to the separatrix (Figure 5.2, p-value: 0.16, r^2: 0.026). Additionally, no significant correlations were found between the RSR and individual expression levels for XIAP or BFAR (Figure 5.5). There was a significant correlation between RSR and expression levels of CASP8 (p-value: 0.049, r^2: 0.052), however this was only due to an outlier with very high expression levels and low RSR. When this individual was excluded, the correlation disappeared (data not shown). There was also a correlation between RSR and the expression level of CASP3 (p-value: 0.045, r^2: 0.053)), which could not be linked as easily to a single

outlier. Interestingly, there was a difference in how much differences in expression of each gene contributed to differences in the calculated time to apoptosis, where XIAP had the strongest correlation between expression level and time to apoptosis (Figure 5.4, p-value: 1.2×10^{28}, r^2: 0.81), followed by BFAR (p-value: 5.6×10^{-6}, r^2: 0.25). Although CASP3 did correlate (p-value: 0.004, r^2: 0.109), the r^2 was much lower than the r^2 for XIAP and BFAR. There was no clear correlation between the expression of either CASP8 and the time to apoptosis (p-value: 0.52, r^2: 0.006).

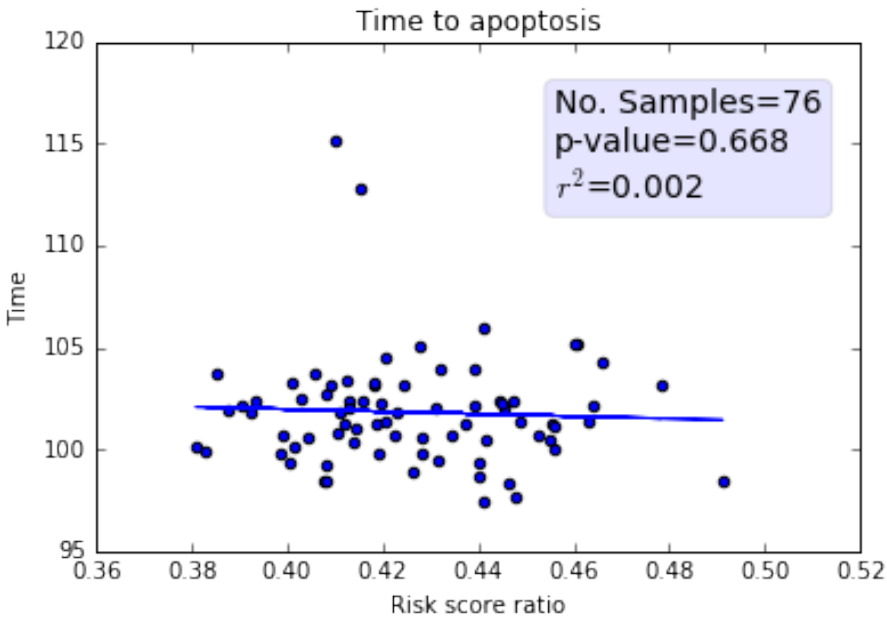


Figure 5.1: Correlation between RSR and time to apoptosis as calculated by the smaller apoptosis model for lymphoblastoid cell lines from the 1000 genome project.

By closer examination, 2 individuals were identified, with a much longer time to apoptosis than the rest of the population. These samples were also found to have a much higher expression of XIAP, compared to the other samples (Figure 5.3). In addition to the 2 female individuals identified, 5 male individuals also had an overexpression of XIAP and a corresponding increased time to apoptosis (data not shown). Examining the exome and genome sequencing data of the abnormal samples showed that there were no differences in copy number variations compared to other samples (Appendix Figure I.15 and I.16). Since the abnormal behaviour was overrepresented in male individuals and the lymphoblastoid cell lines used, had been immortalised through EBV transformation, a method which is known to sometimes affect the expression of XIAP [114, 115], it was hypothesised that these expression values were not representative of the normal behaviour of these samples. It was further hypothesised that removing them from the analysis would result in

a more accurate picture of any effect the SNPs would have on the expression levels and the system behaviour. However, removing these samples did not reveal any correlations between RSR and the time to apoptosis (Figure 5.5, p-value: 0.815, r^2: 0.001), distance to separatrix (Figure 5.6, p-value: 0.14, r^2: 0.03) or the individual expression levels (Figure 5.7) Although, at a FDR of 0.25 the correlation between RSR and expression levels of Caspase 3 was significant.

Twelve SNPs were selected, which through literature research were hypothesised to be affecting at least one of the genes in the model and the same analysis was performed using RSRs of only those SNPs (Table 5.1). This time, there was a significant correlation between RSR and time to apoptosis, both when using all 76 individuals (Figure 5.8, p-value: 0.016, r^2: 0.076) and when excluding the 2 individuals with abnormal response times (Appendix Figure I.2, p-value: 0.008, r^2: 0.094). This trend was negative, meaning that increased RSR was linked to a faster response time to apoptosis. However, there was no significant correlation between RSR and distance to separatrix in either case (p-value: 0.33 and 0.27).

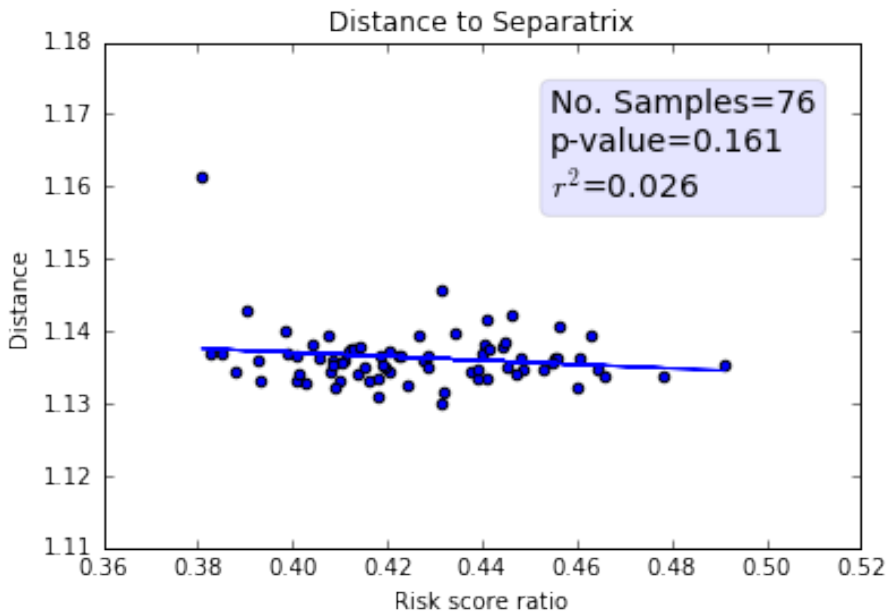


Figure 5.2: Correlation between RSR and distance to separatrix for the lymphoblastoid cell lines from the 1000 genome project.

Out of the 12 SNPs previously chosen, genotype data was available for 9 for the breast tissue later used in this chapter. To be able to compare the results between the two models, the same analysis was performed using RSRs of only those 9 SNPs (Table 5.1). In this case too, there was a significant correlation between RSR and time to apoptosis

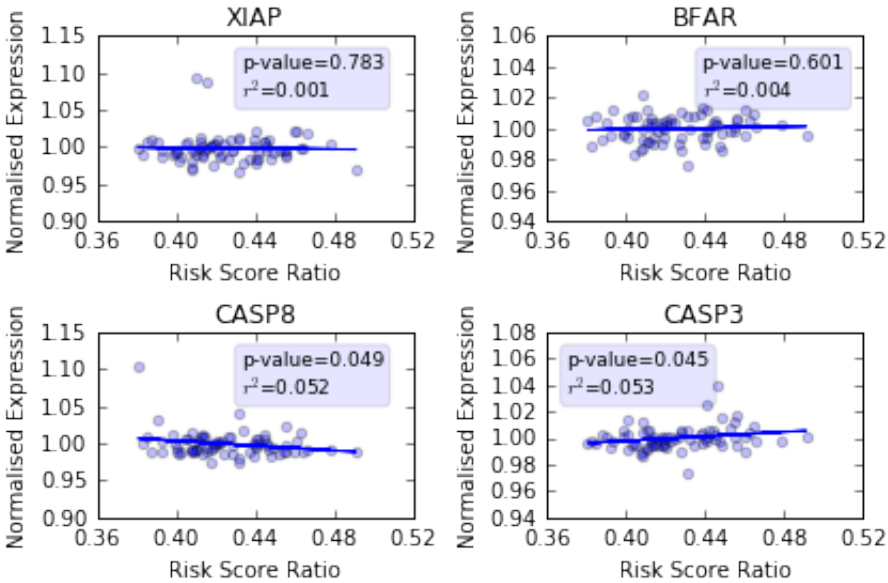


Figure 5.3: Correlation between RSR and expression value in lymphoblastoid cell lines for the four genes XIAP, BFAR, CASP8 and CASP3 (from top left to bottom right).

both when including all individuals (Appendix Figure I.3 , p-value: 0.016, r^2: 0.076) and when excluding the two with an abnormal response time (Appendix Figure I.4, p-value: 0.005, r^2: 0.104). As when using the 12 SNPs, these correlations were also negative, with decreased response time with increased RSR. There was no correlation between RSR and distance to separatrix (p-value, larger set: 0.56; p-value, smaller set: 0.50).

Lastly, 4 out of the previously selected 9 SNPs were selected based on them being transcription regulators with binding sites in a promoter of one of the the genes in the model. As was the case with the previous 2 sets of SNPs, with these 4 SNPs, there was also a negative correlation when excluding the 2 individuals with abnormal response time (Appendix Figure I.5, p-value: 0.022, r^2: 0.070). However, there was no significant correlation when these 2 individuals were included (Appendix Figure I.6, p-value: 0.288, r^2: 0.015). In both cases there was no correlation between RSR and distance to separatrix (p-value, larger set: 0.241; p-value, smaller set: 0.224).

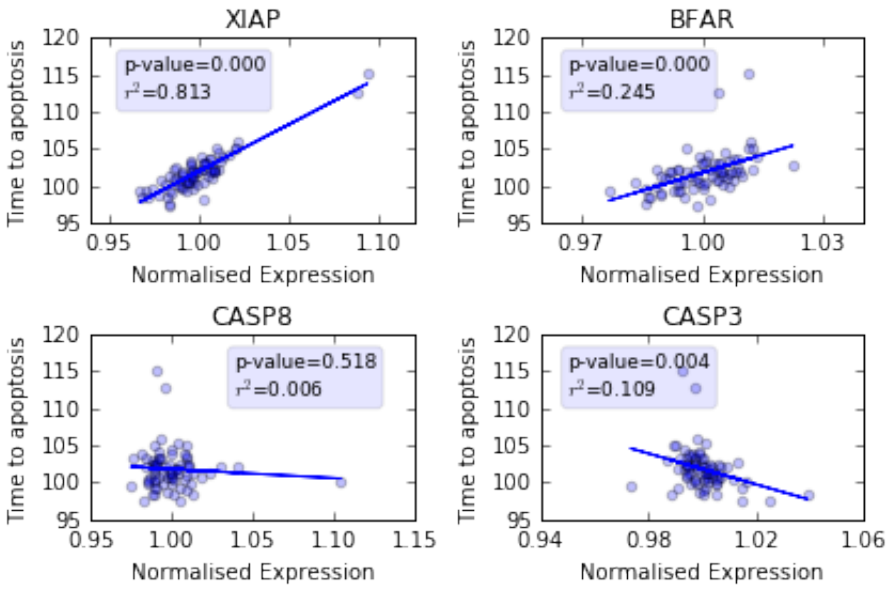


Figure 5.4: Correlation between time to apoptosis and expression values in lymphoblastoid cell lines for the four genes XIAP, BFAR, CASP8 and CASP3 (from top left to bottom right).

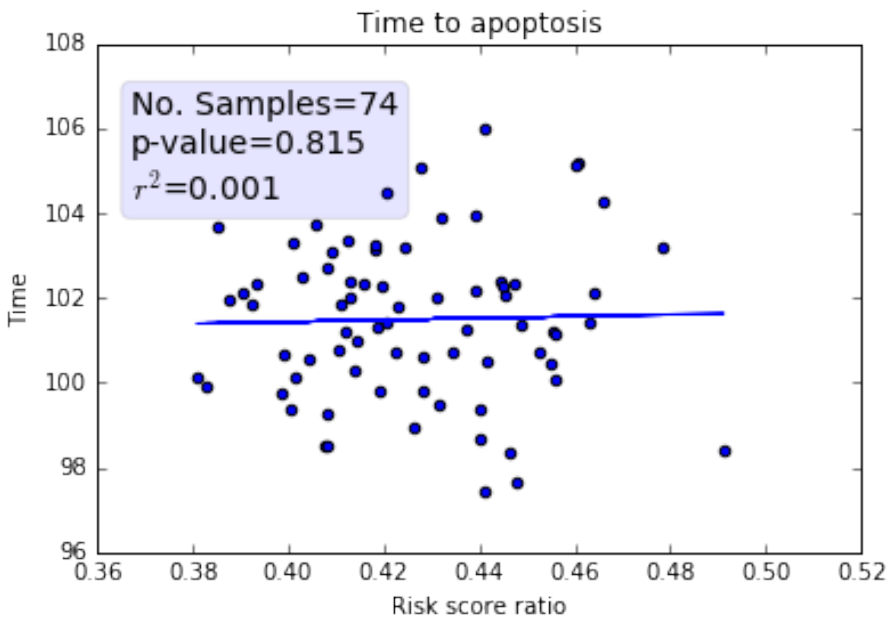


Figure 5.5: Correlation between RSR and time to apoptosis for the lymphoblastoid cell lines from the 1000 genome project, excluding the samples identified in Figure 5.1 as forming a separate cluster of longer response times.

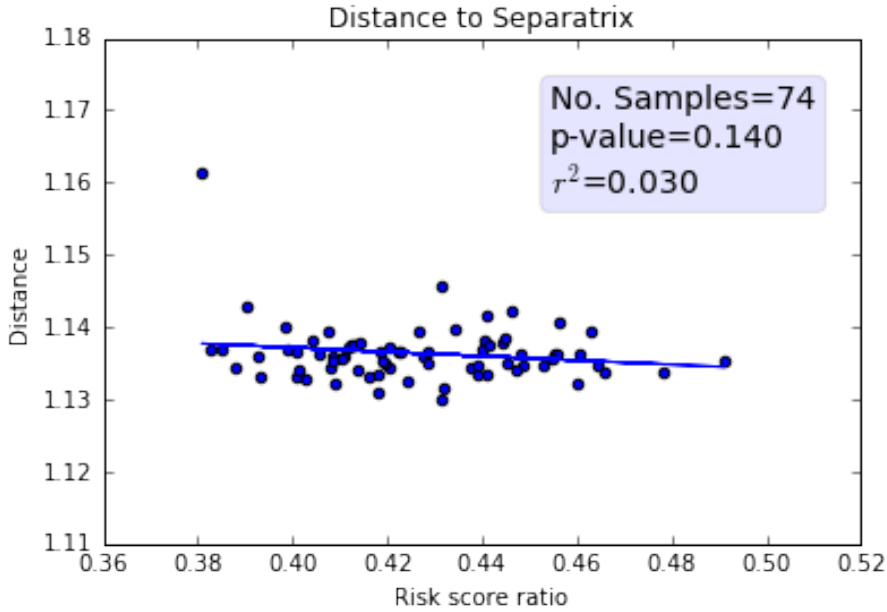


Figure 5.6: Correlation between RSR and distance to separatrix for the lymphoblastoid cell lines from the 1000 genome project, excluding the samples identified in Figure 5.1 as forming a separate cluster of longer response times.

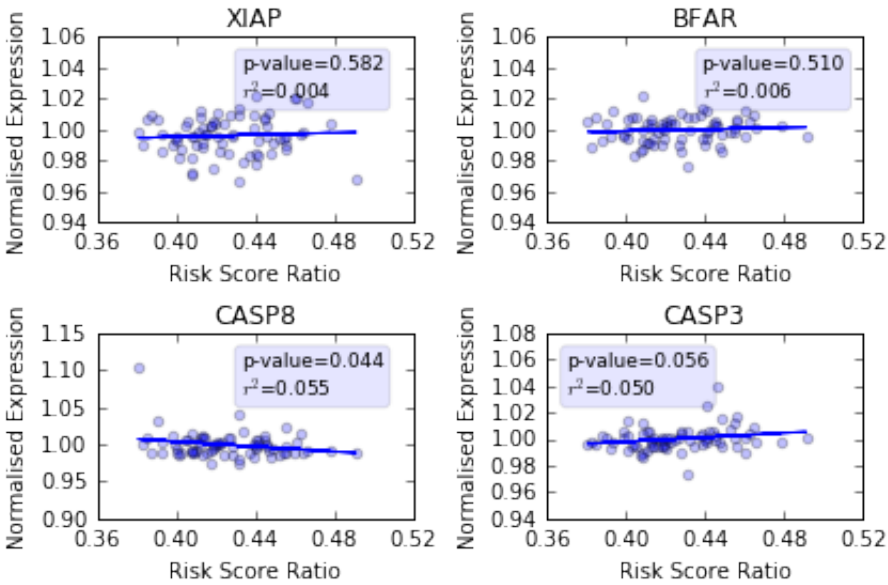


Figure 5.7: Correlation between RSR and expression values for the four genes XIAP, BFAR, CASP8 and CASP3 of lymphoblastoid cell lines form the 1000 genome project, excluding samples forming a separate cluster of behaviour in Figure 5.1.

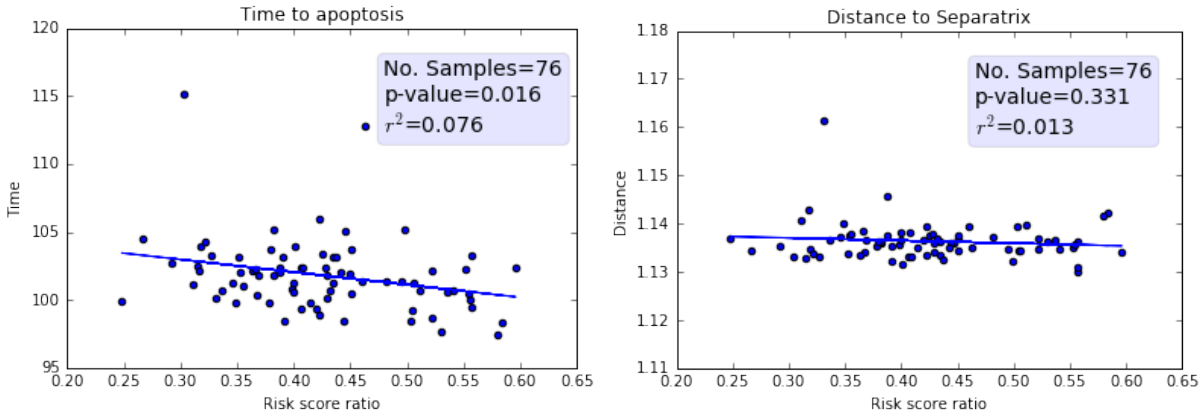


Figure 5.8: Left: Correlation between risk score ratio and time to apoptosis, as calculated by the smaller apoptosis model, for the lymphoblastoid cell lines from the 1000 genome project. Right: Correlation between risk score ratio and distance to separatrix. In both cases 12 SNPs thought to be linked to one of the genes in the model were used to calculate the RSRs.

5.3.2 Breast cancer associated SNPs show weak correlations with model behaviour for breast tissue data from TCGA

In the TCGA database, 85 white women with breast cancer were identified which had genotyping data as well as RNA sequencing data available for normal breast tissue. By expanding the data set of SNPs to include all SNPs within 50,000 bp of the published SNPs associated with breast cancer and with an LD above 0.8, data for 90 SNPs could be extracted for the normal tissue samples.

When normalising the RNA expression values for the four genes XIAP, BFAR, CASP8 and CASP3 around the mean and reducing the variance, the expression of XIAP had a normal distribution according to the Shapiro-Wilk test for normality [113] (p-value: 0.38), The distributions of BFAR, CASP8 and CASP3 were all skewed (p-values: BFAR, 0.009; CASP8, 2.610^4; CASP3, 3.110^4) (Appendix Figure I.7). Using these expression values as input for the smaller apoptosis model the time to apoptosis was calculated and correlated with the RSR including the 90 SNPs extracted previously. Contrary to what would be expected, a small, negative, although not significant (p-value: 0.36), correlation was observed, with shorter time to apoptosis as RSR increased (Figure 5.9).

By filtering out a subset of 9 SNPs (Table 5.1) thought to be affecting genes which in turn could affect the expression of proteins in the model, a positive correlation was identified with a p-value of 0.042 (Figure 5.10).

By randomly selecting 9 SNPs from the total set of SNPs and performing the analysis as previously described, a distribution of p-values for the correlations was attained. Even though this distribution had a slight skew towards negative correlation, the 9 SNPs previously chosen yielded a correlation which was both one of the most positive and had among the lowest p-values possible given the data (Figure 5.11).

When performing the same analysis, using only one SNP at a time, the distribution of the chosen SNPs was skewed towards positive correlations. However, 2 out of the 9 SNPs had a close to zero correlation and one SNP had the most negative, although not most significant, correlation, of all possible correlations, given the entire dataset (Figure 5.12). Furthermore, only one of the 9 SNPs had a significant correlation with time to apoptosis.

When comparing the RSRs directly to the RNA expression values of the four genes in the model, no significant correlation could be identified, indicating that it is not just one, or a few SNPs, targeting one node in the network, driving the change in response time, but a concert of small changes all over the system, which play together to yield the observed correlation (Appendix Figure I.8). This was true, both when considering all of the SNPs in the data set or when only the 9 previously chosen SNPs, as well as when one of the 9 SNPs at a time were used (Appendix Figure I.8-I.11). Although, at a

Table 5.1: SNPs identified through literature research as being likely to affect the expression of any of the genes in the small apoptosis model. The subset of SNPs which could be used for the breast tissue from TCGA has been marked in the forth column. A further selection was made on SNPs mapping to genes which acted as transcriptional regulators and had binding sites in promoters of at least one of the genes in the model. Two SNPs mapped to genes thought to regulate all four genes, whereas two additional SNPs mapped to a gene thought to regulate only BFAR and CASP8.

SNP ID	Nearest Gene	Binding Site for Gene Promoter	TCGA
rs3757322	ESR1	BFAR, CASP8	Yes
rs9397437	ESR1	BFAR, CASP8	Yes
rs11780156	MYC	XIAP, BFAR, CASP8, CASP3	Yes
rs2823093	NRIP1	-	Yes
rs6596100	HSPA4	-	Yes
rs79724016	HIVEP3	-	Yes
rs10760444	LMX1B	-	Yes
rs6569648	L3MBTL3	-	Yes
rs2965183	GATAD2A, MIR640	XIAP, BFAR, CASP8, CASP3	Yes
rs1830298	CASP8, ALS2CR12	-	No
rs2747652	ESR1	BFAR, CASP8	No
rs17156577	CREB5	-	No

FDR of 0.25 the correlation between RSR and expression levels of BFAR was significant (Appendix Figure I.9).

When analysing the relation between the RSR and the distance to the separatrix, no significant correlation was identified, using either all of the SNPs, or the subset of 9 SNPs (Figure 5.13), although the direction of the trend using the subset was in the expected direction. Furthermore, when comparing the results from the regression analysis of the 9 selected SNPs to the distribution of possible correlations of 9 SNPs(Figure 5.14) the correlation of the selected SNPs was far from being among the most significant ones possible (in terms of significance or trend strength). When performing the same analysis for single SNPs there was still a skew of the selected SNPs in the expected direction, but it was not as clear as when looking at the time to apoptosis (Figure 5.15). Altogether, the results indicate that, although the risk score could be correlated with the time to apoptosis, at least for the subset of SNPs likely to be linked to the model, this trend could not be captured by the distance to the separatrix, given the current data.

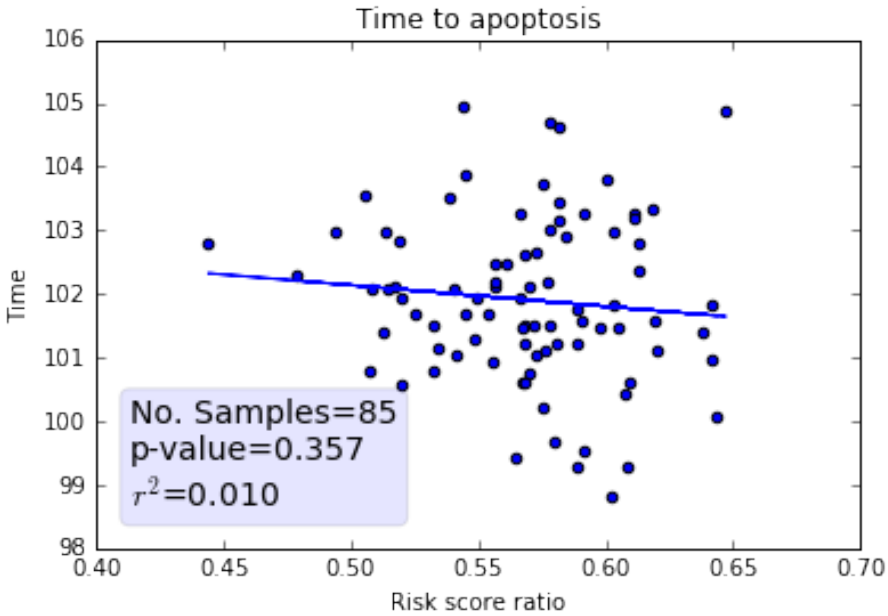


Figure 5.9: Correlation between RSR and time to apoptosis as calculated by the smaller apoptosis model, using 10,000 molecules of activated Caspase 8 as input for normal breast tissue. All 90 SNPs associated with risk of developing breast cancer were used to calculate the RSR.

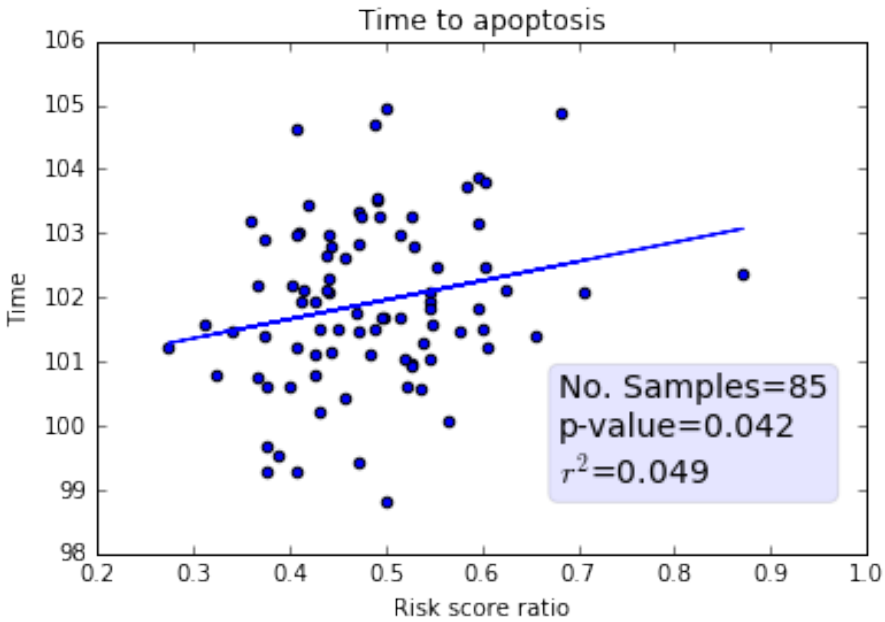


Figure 5.10: Correlation between RSR and time to apoptosis as calculated by the smaller apoptosis model, for normal breast tissue from TCGA. 9 SNPs thought to be linked to any of the genes in the model were used to calculate the RSR and the model was run as previously described.

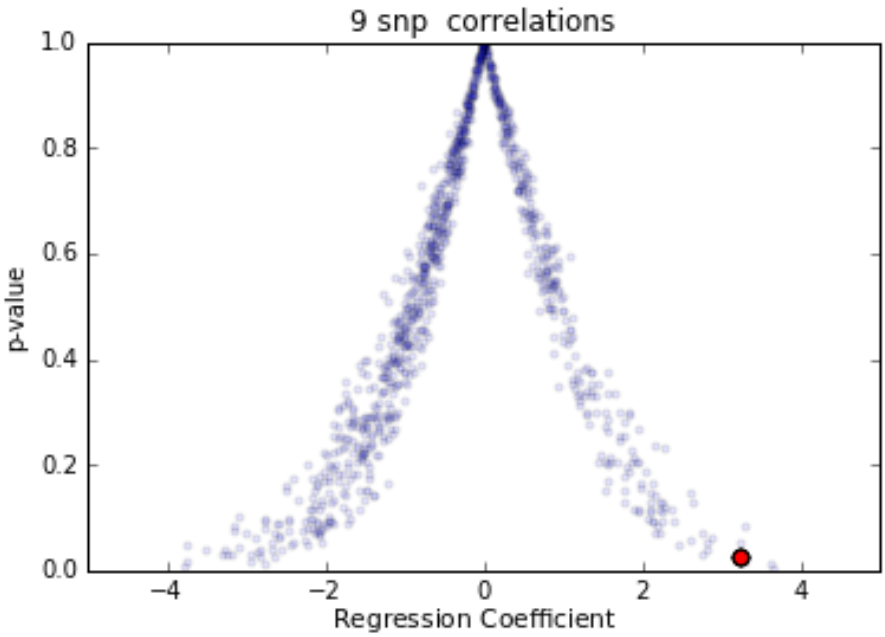


Figure 5.11: Distribution of p-values and regression coefficients for correlations between RSR and time to apoptosis using data for a subset of SNPs in normal breast tissue. 9 SNPs were randomly selected from the breast cancer associated SNP data set and compared to the results from the set of 9 SNPs manually selected previously for further analysis (Red). Using the manually selected SNPs resulted in one of the smallest p-values and one of the largest positive regression coefficients possible given the data it came from.

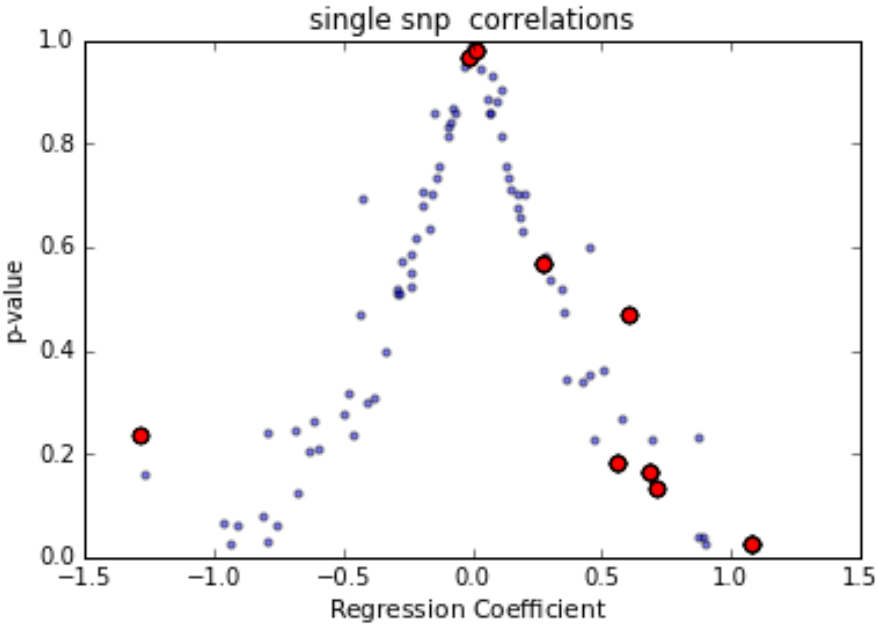


Figure 5.12: Distribution of p-values for correlations between RSR and time to apoptosis for normal breast tissue. Single SNPs were used to calculate the RSR. The 9 SNPs chosen previously for further analysis (red) had a distribution of regression coefficients which was skewed to upper values of the total distribution given all the data.

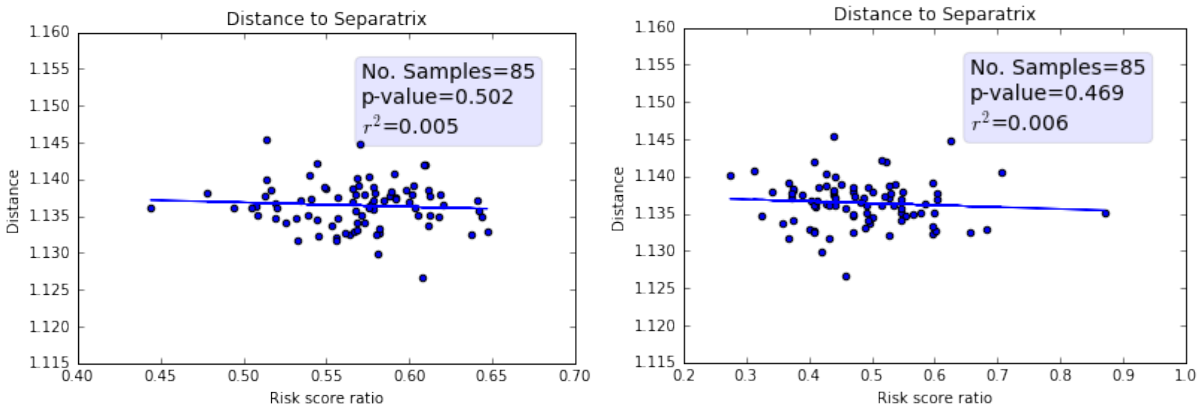


Figure 5.13: Correlations between RSR and distance to separatrix surface for the smaller apoptosis model for normal breast tissue. Left: all SNPs in the breast cancer SNP data set were used. Right: only the 9 previously selected SNPs were used to calculate the risk score ratio.

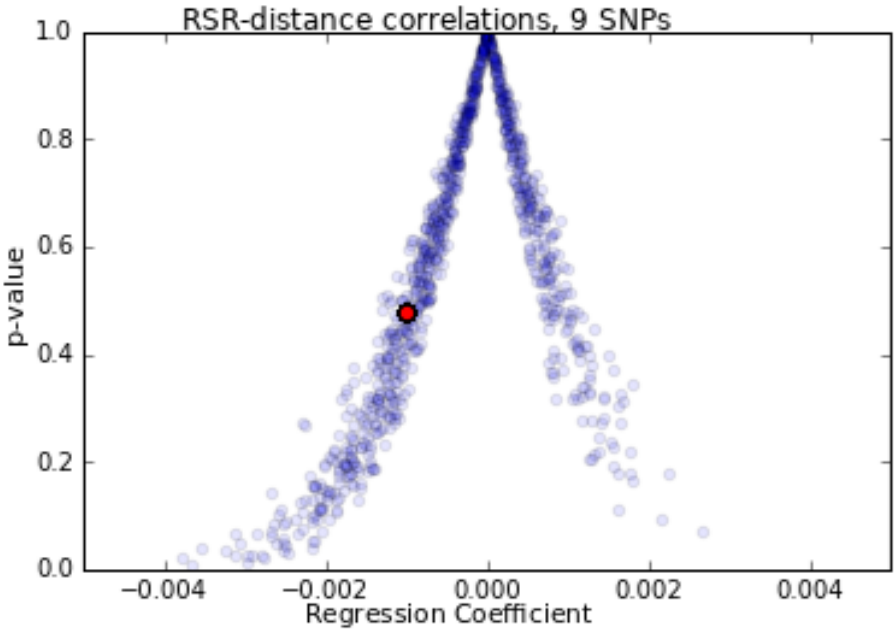


Figure 5.14: Distribution of p-values and regression coefficients for correlations between RSRs and distance to separatrix surface for the smaller apoptosis model using SNP data from normal breast tissue. Using 9 SNPs randomly selected from the total data set of SNPs associated with breast cancer resulted in a distribution of regression coefficients slightly skewed towards negative values. Usign the set of 9 SNPs previously selected for further analysis resulted in a regression coefficient slightly to the below the centre of the distribution.

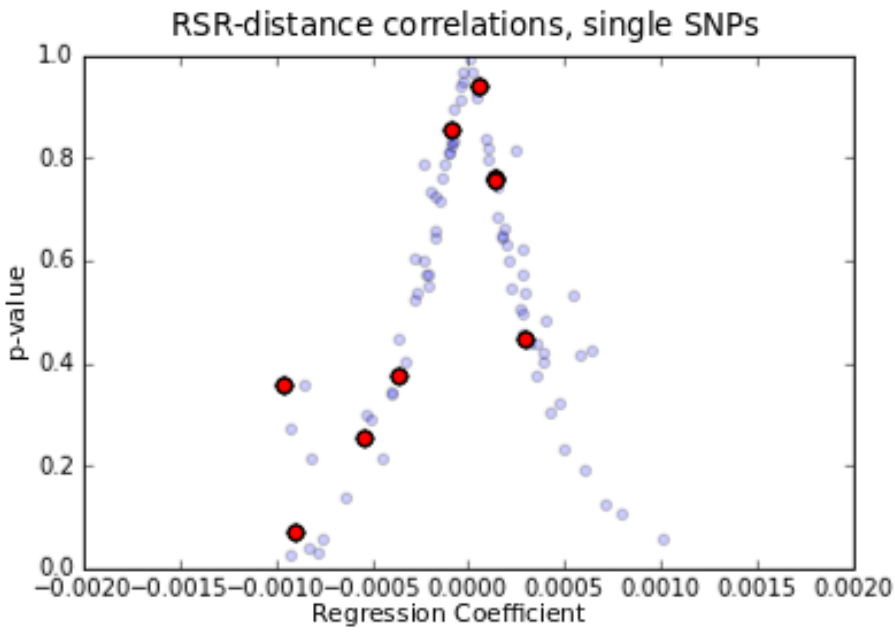


Figure 5.15: Distribution of p-values and regression coefficients for correlations between RSR, using one single SNP at a time, and distance to separatrix surface for the smaller apoptosis model for normal breast tissue. The red dots correspond to the SNPs previously selected for further analysis.

Table 5.2: Prostate cancer associated SNPs identified as being likely to affect the expression of any of the genes in the small apoptosis model. One of the SNPs is close to CASP8 which is itself in the model. The other SNPs are close to transcription regulators which have binding sites in the promoters of at least one of the genes in the model.

SNP ID	Gene Name	Binding Site for genes
rs6062509	ZGPAT	CASP3, CASP8, XIAP, BFAR
rs7094871	TCF7L2	CASP3, CASP8, XIAP, BFAR
rs11290954	C11orf30/EMSY	CASP3, CASP8, XIAP, BFAR
rs4962416	CTBP2	CASP3, BFAR
rs10460109	TSHZ1	XIAP, BFAR
rs11480453	DNMT3B	CASP3, CASP8, BFAR
rs59308963	CASP8	-

5.3.3 Risk scores for prostate cancer associated SNPs do not correlate with time to apoptosis or distance to separatrix

From the paper published by Schumacher *et al.* in nature Genetics 2018 [32], 142 SNPs were identified to be associated with risk of developing prostate cancer. Out of these SNPs genotypes for 86 SNPs could be extracted (either directly or on the basis of LD as described in Section 5.1.1) for 50 normal prostate tissues from the TCGA database, which also had expression data for the four genes in the model, XIAP, BFAR, CASP8 and CASP3. After normalising the RNA expression data, only BFAR had a normal distribution of expression, whereas XIAP, CASP8 and CASP3 where skewed according to the Shapiro-Wilk test for normality [113] (p-values: XIAP, 0.04; BFAR, 0.25; CASP8, 0.003; CASP3, 0.002) (Appendix Figure I.12). Using all SNPs to calculate the RSR for each sample, the correlation between the RSR and the output of the model, time to apoptosis, was evaluated. Similar to the breast cancer data, there was a (negative) non-significant correlation, contrary to what would be expected if the SNPs did indeed affect the genes in the model (Figure 5.16).

When examining the function of the genes closest to the SNPs, 7 SNPs were identified to be either affecting the gene directly, or affecting transcription regulators with binding sites in the promoter region of any of the genes (Table 5.2).

Using only these 7 SNPs to calculate the RSR a positive correlation can be seen, although the p-value was not significant at 0.103 (Figure 5.17). Randomly selecting 7 SNPs multiple times and performing the same type of correlation analysis results in a distribution centred around a slope of zero and p-value of 1 (Figure 5.18). The set of 7 SNPs chosen for further analysis yields a correlation which is positioned at the upper end of this distribution. However it is not among the very best correlations.

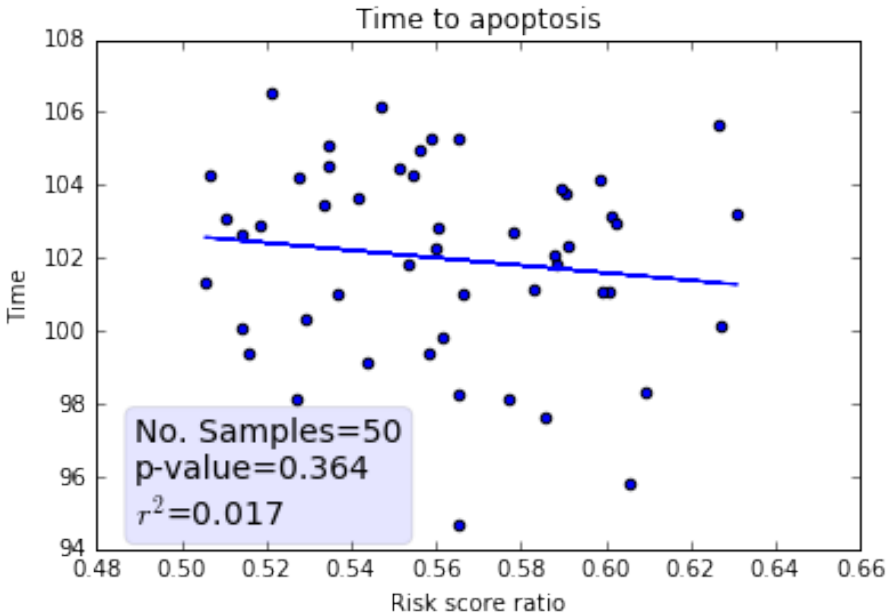


Figure 5.16: Correlation between RSR and time to apoptosis as calculated by the smaller apoptosis model for normal prostate tissue, using 10,000 molecules of activated Caspase 8 as input. All 86 SNPs associated with risk of developing prostate cancer were used to calculated the risk score.

Using only one SNP at a time and performing the correlation analysis yields a distribution similar to that when using 7 SNPs (Figure 5.19). Although, 6 out of the 7 SNPs had a positive correlation with the time to apoptosis, none was significant. However, the clear difference in distribution compared with the total set of SNPs suggest that at least some of the SNPs do indeed have an effect on the genes in the model.

A similar trend can be seen when comparing the RSR to the expression values of 3 out of the 4 genes in the model (Appendix Figure I.13 and I.14). When using 7 SNPs there was a positive correlation with XIAP and negative correlations with CASP8 and CASP3, although all were non-significant. When looking at one SNP at a time there was a skew towards positive correlations for XIAP and towards negative correlations with CASP8 and CASP3. However, for BFAR, the trend was towards negative correlations, (both with 1 and 7 SNPs), contrary to what would be expected given that BFAR acts as an inhibitor for apoptosis.

When comparing the cumulative RSR of the 7 SNPs and the distance to the separatrix of the model no significant correlation was found (Figure 5.20). However, like with the analysis of the RSR and the time to apoptosis, the correlation was in the expected direction and the distribution of possible correlations given the entire data set was weighed towards the opposite direction. When looking at individual SNPs the distribution of correlations

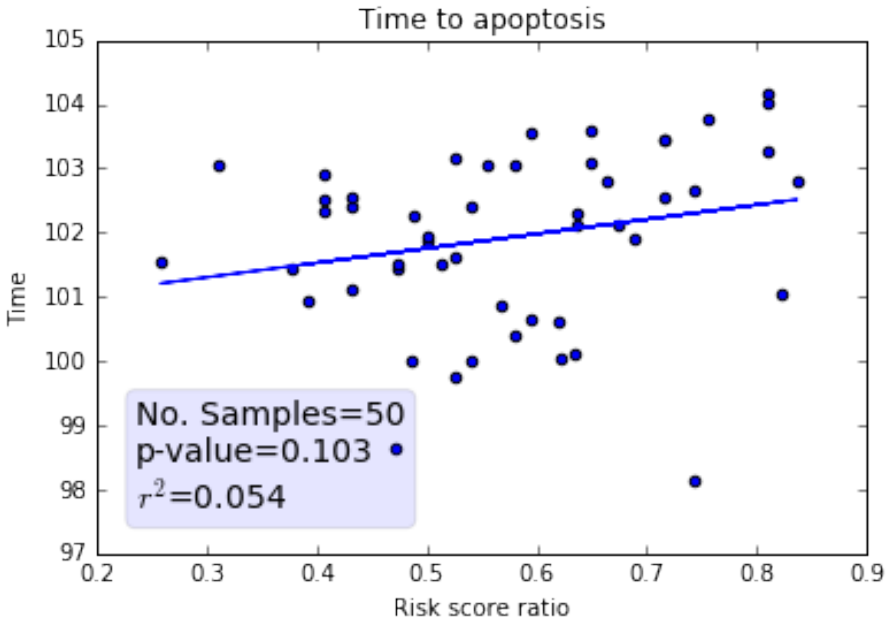


Figure 5.17: Correlation between RSR and time to apoptosis for 50 prostate tissues using 7 SNPs likely to affect the expression of any of the genes in the model.

for the selected SNPs was also skewed in the expected direction, although none of them were significant (Figure 5.21).

As a negative control for both types of cancer, the breast cancer associated SNPs were analysed on prostate tissue and the prostate cancer associated SNPs were analysed on breast tissue. In both cases there was no significant correlation between the RSR and the time to apoptosis (Figure 5.22 and 5.23). This was true, both when using the entire data sets or when only using the 9 breast cancer associated or 7 prostate cancer associated SNPs.

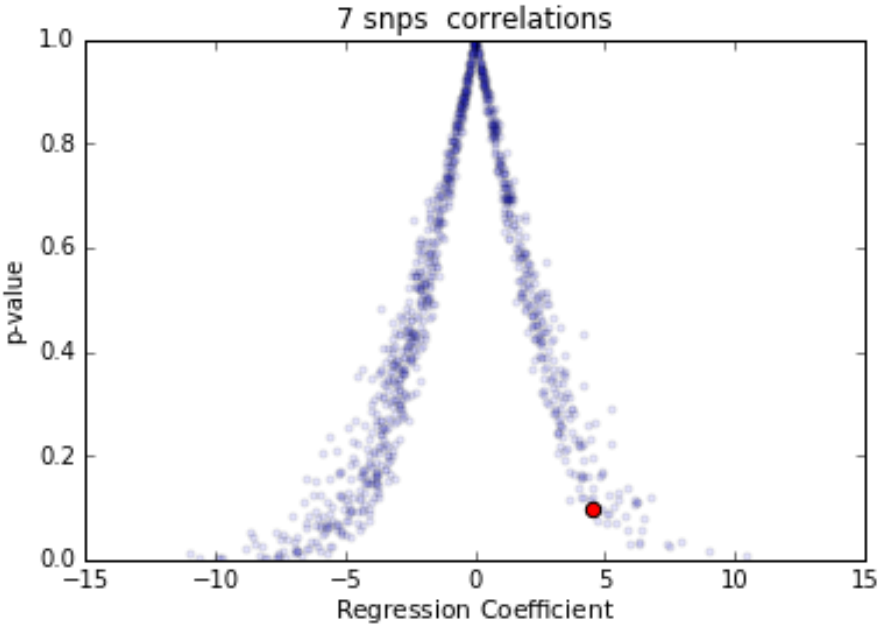


Figure 5.18: Distribution of p-values and regression coefficients for correlations between RSR for 7 randomly selected SNPs and time to apoptosis as calculated by the smaller apoptosis model for normal prostate tissue. Using the set of 7 SNPs previously selected for further analysis (red) resulted in a regression coefficient located in the upper region of the distribution, although not at the very top.

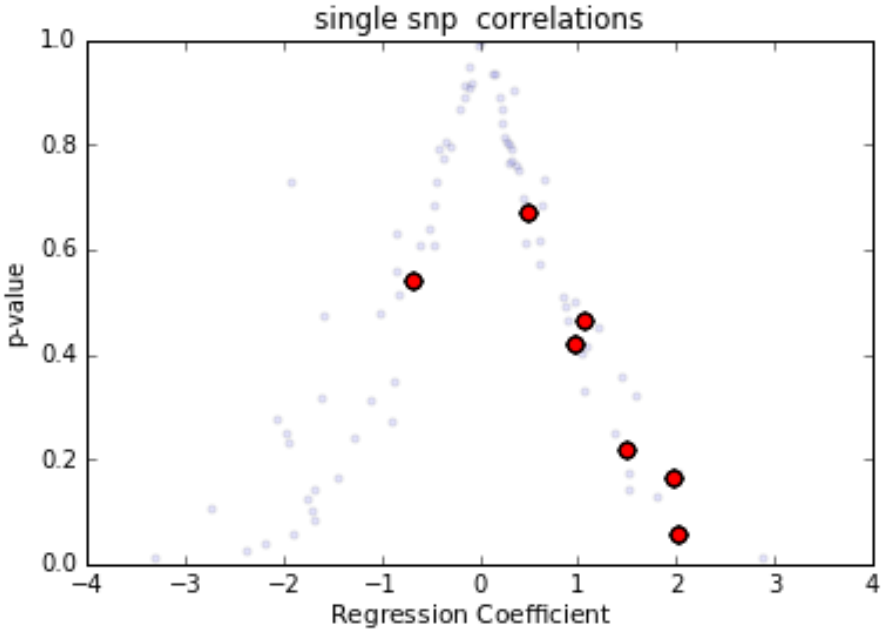


Figure 5.19: Distribution of p-values and regression coefficients for correlations between RSR for single SNPs and time to apoptosis as calculated by the smaller apoptosis model for normal prostate tissue. Using the 7 previously selected SNPs (red) resulting in a distribution skewed towards upper values of the total distribution given all the data.

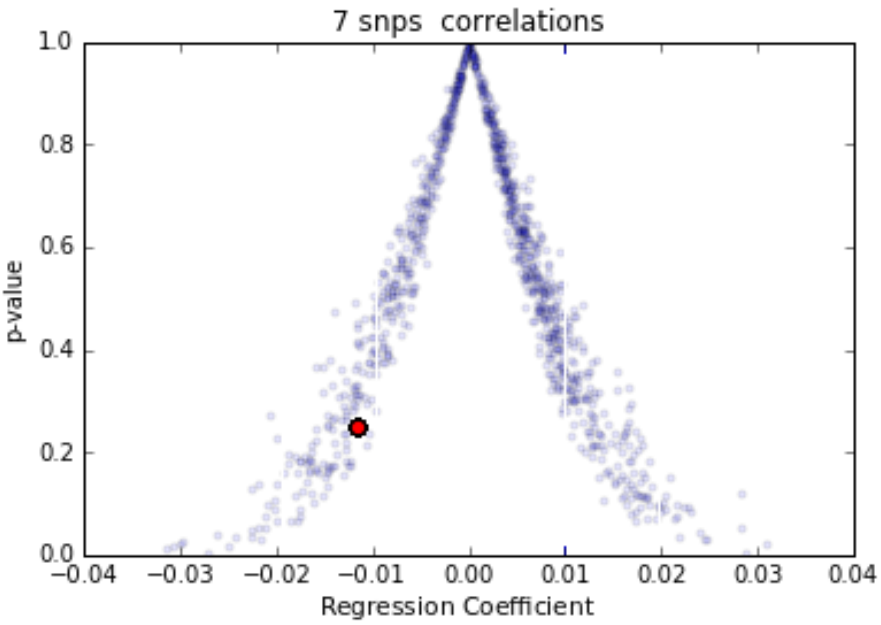


Figure 5.20: Distribution of p-values and regression coefficients for correlations between RSR of 7 randomly selected SNPs and the distance to separatrix surface for the smaller apoptosis model for normal prostate tissue. Using the 7 SNPs previously selected for further study (red) resulted in a regression coefficient far to the left of the total distribution given all the data.

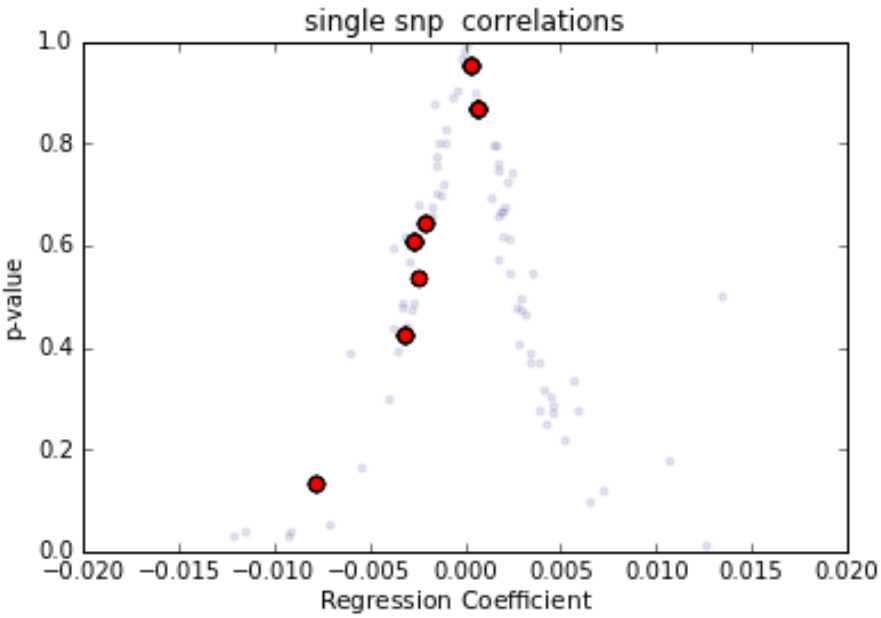


Figure 5.21: Distribution of p-values and regression coefficients for correlations between the RSR of a single SNP and the distance to the separatrix for the smaller apoptosis model for normal prostate tissue. The 7 SNPs previously selected are marked red.

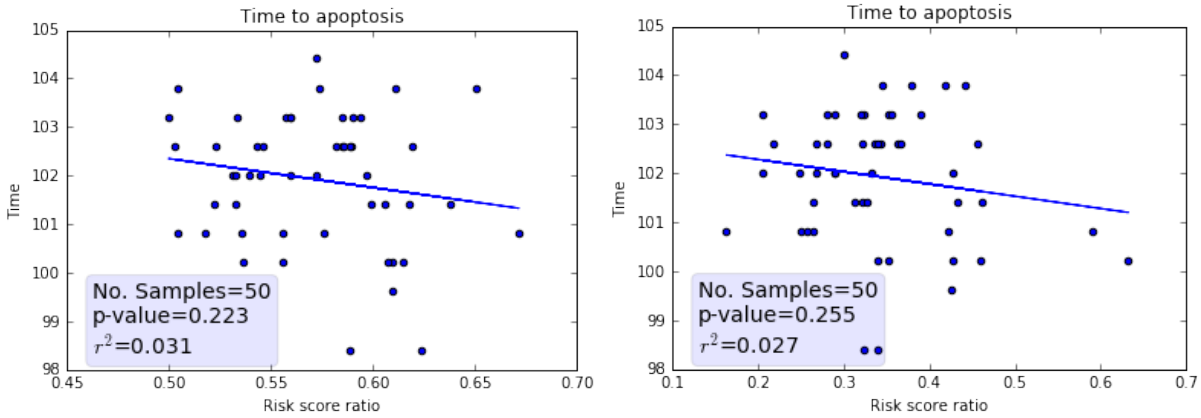


Figure 5.22: Correlation between RSR and time to apoptosis for 50 prostate tissues using breast cancer associated SNPs. Left: All breast cancer associated SNPs were used. Right: The 9 previously identified breast cancer associated SNPs were used.

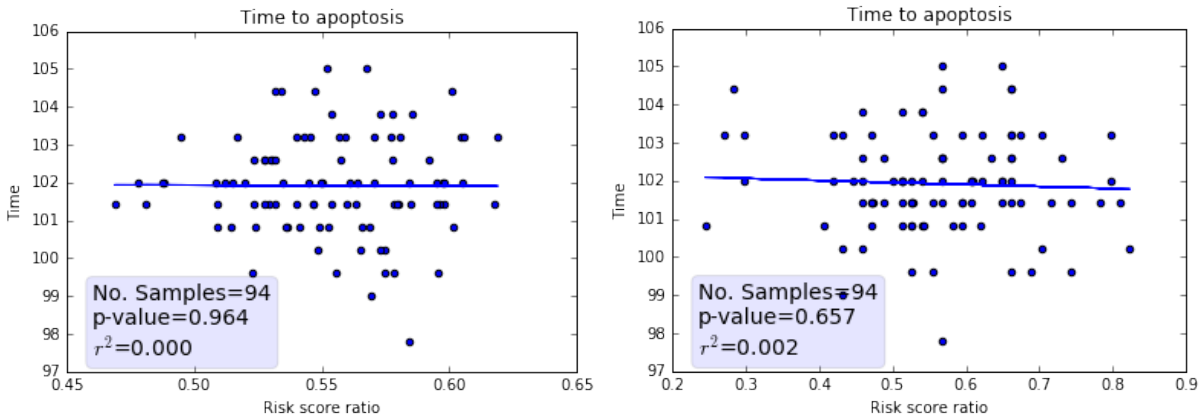


Figure 5.23: Correlation between RSR and time to apoptosis for breast tissues using prostate cancer associated SNPs. Left: All prostate cancer associated SNPs were used. Right: The 7 previously identified prostate cancer associated SNPs were used.

5.4 Discussion

Using simulated data in Chapter 4 the link between the risk score and the risk of developing a phenotype was assessed and verified. The data suggested a clear trend between RSR and time to apoptosis. Furthermore, the separatrix analysis method was shown to be able to capture the increased risk of developing a cancer phenotype by taking into account the effect the risk associated SNPs had on the parameters in the model and how that shifted the average distance of the individual to the separatrix surface. In this chapter the same method was applied to experimental data from three different sources, lymphoblastoid cell lines, from the 1000 genome project, and normal tissue of breast and prostate from TCGA.

First, lymphoblastoid cell lines from the 1000 genomes project were used as an experimental model to assess the validity of the results from Chapter 4. Unfortunately, no correlation was found between the RSR of breast cancer associated SNPs and the time to apoptosis, extracted from the model dynamics, or distance to separatrix. However, there was a clear correlation between the time to apoptosis and the expression values of XIAP and BFAR and a weaker correlation between the model output and the expression values of CASP3 (Figure 5.4). If parts of this trend could be linked to a subset of the risk associated SNPs, then the RSR using only those SNPs could potentially be correlated with the output of the model and the distance to separatrix.

Furthermore, from the analysis in Chapter 4, it could be seen that, although only a subset of the SNPs in the analysis has to be linked to the model in order to capture the correlation, especially when analysing the distance to separatrix, this would have to be a substantial part of the total set of SNPs. In fact, when there were only 50 samples, around 70% of the RSR had to be linked to the model for the simulations to yield a significant correlation (p-value < 0.05) in at least 95% of the cases (Figure 4.26a). However, when using 100 samples a little less than 50% of the RSR had to be linked to the model for the same results (Figure 4.26b).

Given the small number of genes in the model of investigation, the number of SNPs affecting the expression of any of these genes in the total SNP data set is not likely to be very large. Indeed, when looking at the genes closest to the SNPs, not a single SNP was linked to any of the genes in the model. However, through literature research, 12 SNPs were identified which were deemed likely to affect the expression of at least one of the four genes, either by being transcriptional regulators with a binding site in the gene's promoter, or by otherwise having been associated with gene expression regulation. Out of these 12 SNPs, the genotypes for 9 SNPs were available for the breast tissue later

used in this chapter. Furthermore, out of these 9 SNPs, 4 SNPs were indicated to be transcriptional regulators with binding sites in one of the genes promoters. Using the set of 12 and 9 SNPs, there was a significant correlation between the RSR and time to apoptosis, but not using only the set of 4 SNPs. Interestingly the trends in all 3 cases were towards decreased time to apoptosis as RSR increased.

It is known that the immortalisation process can affect the expression of XIAP [114, 115] and in the total data set 7 individuals were identified (2 female and 5 male), which had a much higher expression of XIAP. These individuals also had a much longer time to apoptosis as calculated by the model. Upon examination, these individuals did not have any copy number aberrations in the region around the XIAP gene and the difference in XIAP expression was attributed to the immortalisation process. Excluding these samples did result in a significant correlation using the smallest subset of 4 SNPs as well. It did not, however, alter the direction of the trends in any of the cases.

This might be contrary to what would be expected, given that the increased sensitivity resulting from decreased response time would be expected to kill cells before they develop to cancer cells. However, this hypothesis is only valid for tissues for which the SNP is associated with cancer susceptibility. It is possible that a SNP associated with breast cancer is acting by desensitising the immune system and decreasing the immune response to the breast cancer cells. In fact, several breast cancer associated SNPs have been predicted to alter the expression of genes known to play a role in immune system related cells [116]. Furthermore, these correlations were not seen using male individuals, suggesting that the associations are gender specific.

Even if these results are not what would, at first, be expected, they are not directly inconsistent with the hypothesis, that the SNPs act together to affect the dynamics of the model and thereby increase the risk of cancer developing. However, it does not verify this effect in the target tissue. The focus was therefore shifted to breast tissue samples from TCGA.

Using all SNPs associated with breast cancer, no correlation between RSR and either time to apoptosis or distance to separatrix was found. However, when the SNPs were limited to the 9 SNPs previously identified to be likely to affect any of the genes in the model a significant correlation between RSR and time to apoptosis was found (p-value: 0.042). Furthermore, when comparing this correlation with all possible correlations of 9 SNPs from the original data set, it proved to be one of the most positive and with one of the lowest p-values. When comparing the correlations using a single SNP at a time, the set of 9 SNPs chosen was overrepresented among the more positively correlated SNPs, although, only one was significant. Although this does not prove that there is in fact a

causation between the risk score and the model output, it suggests that there is indeed a link between the two measures.

When comparing the RSRs of either the 9 SNPs or single SNPs at a time, there was no significant correlation between the RSR and the expression values of the four genes in the model. This indicates that the the correlation between the RSR and time to apoptosis is not due to a few SNPs affecting one or two genes in the model, but the result of small perturbations caused by several SNPs over larger parts of the network. Furthermore, it indicates that it is not a one-to-one relation between a SNP and change in expression of a gene.

Given that the correlation between RSR and time to apoptosis was just below significance level, and that the results from Chapter 4 indicate that the distance to separatrix is more weakly correlated with RSR than time to apoptosis, it is not surprising that no significant correlation could be identified, even when limiting the analysis to the 9 selected SNPs. However, even though the correlation using the subset of 9 SNPs was not significant, it was in the right direction, with a shorter distance as RSR increased.

To further assess the possibility of the method to link genotype to risk of phenotype change, the correlation between prostate cancer associated SNPs and time to apoptosis as well as distance to separatrix in normal prostate tissue was investigated. This was a smaller data set than that for breast cancer (50 samples for prostate cancer and 85 for breast cancer), but there is a large number of SNPs associated with prostate cancer development. As with breast cancer, there was no correlation between RSR and time to apoptosis using all SNPs in the data set. When selecting 7 SNPs thought to be affecting any of the genes in the model, a weak correlation was seen, although not significant with a p-value of 0.104. As with breast cancer, the correlation of the smaller data set with time to apoptosis was one of the better correlations possible given the data. Also, when looking at single SNPs at a time, there was a clear preference for positive correlations, compared to the total data set.

Again, as with the breast tissue samples, when looking at the expression values, there were no significant correlations using either all 7 SNPs or one SNP at a time. However, the correlations for the 7 prostate cancer associated SNPs were further from the centre of the distributions than those for 9 breast cancer associated SNPs. Likewise, the single correlations of the 7 SNPs were more skewed towards positive ones for the inhibitors XIAP and BFAR and towards negative ones for the two activators CASP8 and CASP3, than the equivalent correlations for the 9 breast cancer associated SNPs.

Although the RSR could not be linked to the distance to separatrix in any of the three sample sets and it could only be linked to the time to apoptosis in the lymphoblastoid

cell lines and the breast tissue dataset, it is worth pointing out that for both breast and prostate tissue the regression trend did go in the expected direction, that is with a decreased distance with increased RSR. Furthermore, the correlation between RSR and the time to apoptosis for the prostate tissues was in the expected direction (increased time with increased RSR), although not significant. The distributions of the correlations using single SNPs, both when correlating with time to apoptosis and distance to separatrix, were skewed in the expected direction compared to the distribution of the total dataset. Furthermore, the negative controls of both tissues indicate that the trends are tissue specific, as would be expected. Altogether, these results suggest that these methods could indeed be used to link the RSR to the mechanics of increased risk of developing cancer through these dynamical models. What prevented such a link from being statistically verified in this study was likely the sample size.

From the theoretical analysis in Chapter 4 it could be seen that, using 50 samples, a significant correlation between RSR and distance to separatrix could be expected most times, even when very low fractions of the the RSR is linked to the model (Figure 4.26). A correlation with time to apoptosis could be expected to be significant with an even lower fraction of the RSR actually effecting the model and consequently it could be argued that a correlation should have been seen for both tissues (Figure 4.25). However, the theoretical work assumed a perfect correlation between expression values and protein levels and a perfect correlation between genotype and expression values. The experimental data for the two tissue types do have noise in both of these correlations and any correlation between the RSR and the model output would therefore be expected to be much weaker than the theoretical work might suggest.

Since the correlation between RSR and time to apoptosis was just below significance level for breast tissue with 85 samples, it is not surprising that the correlation was not significant for prostate tissue with only 50 samples. Likewise, since the correlation with distance was shown to be weaker in the previous chapter, it is not surprising that there was no significant correlation in either case using the experimental data.

To truly be able to verify the method, a larger data set would have to be used. Unfortunately no such dataset was to be found at the time.

The links between RSR and model output in all three cell types, combined with the successful correlation between RSR and distance to separatrix using simulated data (Chapter 4), do suggest that the work performed is a path worth further exploration. Furthermore, the fact that the RSR barely correlated with any expression values and that no single SNPs correlated with the model output suggests that these type of models are

necessary in order to identify synergetic effects of larger sets of SNPs, which have small effects on larger parts of a pathway.

Chapter 6

Conclusions

The work presented in this thesis was based on a very ambitious project. By combining what is generally known about the biology of the cell and cancer development, with genetics and mathematical modelling, we aimed to gain a deeper understanding of the mechanistics behind the risk of developing cancer. More precisely we were interested in finding out how genetic variations with very small effects on the initial system could cause an increased risk of developing cancer over time.

The introductory chapter in this thesis played two roles; first, to give an understanding of the biology of cancer development and the role genetics plays, as well as to show how extensive the problem of understanding cancer development and the risk of developing cancer is; second, to show various ways in which mathematical models have been used so far in this pursuit, and what possibilities and limitations have been encountered when using them. It became apparent that no dynamical model will ever be able to cover the full scope of cancer development. Instead, previous research has focused on small parts of key pathways in order to gain a deep understanding of some of the mechanisms which drive carcinogenic behaviour.

6.1 Understanding Risk

When trying to understand the way genetics affects the risk of developing cancer there is no one obvious part of the process to focus on. As was seen in Chapter 1, genetic variations associated with and altering the risk of developing cancer can be found spread over large parts of the genome. In Chapter 2 attempts were made to find pathways with relevant enrichments of breast cancer associated Single Nucleotide Polymorphisms (SNPs). The term "relevant enrichment" is important as not all pathways known to be important in cancer development have been studied to the same extent. As a consequence, the degree to which deep knowledge required for building mathematical models has been acquired

and the extent to which the knowledge has been translated into models of value with regards to predictability varies greatly between the pathways.

The choice to map SNPs to the Boolean model of the carcinogenic process was an attempt to anchor any enrichment onto the key genes of the identified pathways, in the hope of being more likely to identify models covering these genes. It quickly became clear that the very limited set of genes in the model was not enough to link the SNPs to models. Even extending the network by considering intermediate connections, thereby allowing the SNP to be two-times-removed from the gene in the model, yielded a surprisingly small number of connections. However, a small number of SNPs could be linked to genes involved in, or thought to be regulating, the apoptotic pathway. This became the main entry point for further studies.

Out of the many models of the apoptotic pathway available, two complementing models were chosen. The first is a large model by Schlatter *et al.* [91] capturing both intracellular and extracellular signalling of apoptosis. The other model, published by Eissing *et al.* [95] is smaller and focuses the core parts of the apoptotic pathway, the activation of Caspase 8 and Caspase 3. While the larger model gives a more complete picture of the apoptotic signalling pathway, the smaller model allowed for a more in-depth analysis of the dynamics due to its lower dimensionality.

In this work, only deterministic models were considered. While it would not have been possible to always use the same tools, the concepts explored in this thesis could have been applied to stochastic models as well. However, this would have introduced more complexity and uncertainty to an already complex question. When investigating the sensitivity of the models, the stochastic results would have made it even more difficult to establish the effect of a SNP on the system outcome. When investigating the distance to the separatrix consideration would have had to be taken to the fact that what appeared as a surface in a deterministic system, would have taken the shape of a density cloud, where each point was associated with some probability of crossing over, making it even harder to define and measure the distance to the separatrix. It is also not clear how such a model could have been evaluated on biological data as was done in Chapter 5, as publicly available expression data of tissues is almost exclusivity bulk sequencing of many cells. This data is more suitable for deterministic systems, which often are designed to capture the general, average behaviour.

6.2 Does Model Sensitivity Explain Risk?

Once a pathway had been identified and two models had been chosen for further study, the next task was to establish a methodology for evaluating model dynamic sensitivity in terms of risk of causing the system to change from a phenotype representing a healthy cell to a phenotype representing a cancer cell. In the case of the two apoptosis models chosen, the signal representing the onset of apoptosis was activation of Caspase 3. Since cancer cells are characterised by evading the onset of apoptosis, the natural criteria for the two phenotypes became whether the concentration of activated Caspase 3 had reached a predefined threshold within the time limit of the simulation.

In Chapter 3 the general theory of model sensitivity was covered and a set of standard sensitivity tools was used to explore areas of parameter or variable space where the model is more sensitive to perturbations caused by genetic variations.

The analysis showed that both models were very robust. Although small gradual changes in the time to onset of apoptosis could sometimes be seen upon modest perturbations of initial concentrations of proteins in the larger model, these changes were almost non-existent in the smaller apoptosis model (upon perturbations in the production rates), especially when stronger initial activation signals were considered. Furthermore, often large changes in initial component concentration (or production rate in the case of the smaller apoptosis model) of a single variables were required to shift the system output from one corresponding to apoptosis to one corresponding to indifference to activation signalling. This could be expected; If the system was operating close to the edge of the basin of healthy phenotype (i.e. the set of system configurations for which the cell would respond to the initial activation signal within the set time limit), cells would be expected to frequently exhibit the aberrant phenotype by chance (due to the stochastic variation between cells). It is also consistent with the results by Schlatter *et al.* [91], showing in their paper that the time dependent behaviour of the Bax-Bak complex, JNK, and Caspase 3 is very robust to most small local changes (increase of 1-10%) in single parameters with Caspase 3 activation being especially robust. In addition to that they also showed that the amount of final active Caspase 3 was very robust towards changes in parameters, where many single parameter changes could span 4 orders of magnitude without changing the phenotypic outcome of any of the tested scenarios of signal activation or drug treatment.

Changing one or a few parameter at a time, however, has its limitations. It is not possible to detect how multiple parameters interact to regulate behaviour, and how a change in one parameter can change the sensitivity of the system with regards to another

parameter. An attempt to overcome this limitation was made with the parameter scans, where two parameters were changed at the same time, showing that the output often did indeed depend on both of them. However, looking at two parameters at a time also meant that the number of analyses increased dramatically. Furthermore, this approach would have been difficult to extend beyond the interaction of three parameters, due to the difficulties in visualising higher-dimensional spaces. Sensitivity analysis using SASSy and SloppyCell revealed that the dynamics of the studied model is not dictated by a few parameters, or even a few low dimensional Principal Components (PCs). Instead, the singular value spectrum indicates that the behaviour of the system is spread over a large number of PCs and even if they decrease in importance, it is not possible to directly set a cut-off for important and unimportant PCs. This phenomenon has been seen in a number of other systems and it has even been suggested that it is a general characteristic for biological systems [102]. Furthermore, by combining SloppyCell with the parameter scan, and assessing the sensitivity of all parameters at different starting positions, a broader understanding could be gained about how the sensitivities change as the system develops from a normal state to a more cancer-like state.

6.3 Development of New Sensitivity Method

Using the sensitivity tools mentioned above, it could be shown that the sensitivity of the system depended on how deep within the basin of healthy phenotype the system was. This behaviour exposed a potential to use these tools to assess under what circumstances genetic germline variations could have an increased effect and cause a disease phenotype to occur; however, it became clear that such an analysis would be very complex and time consuming if larger numbers of the possible mutations an individual could acquire during a lifetime were to be investigated. It would also be very difficult to say anything about the relative effect size of two or more SNPs as the system's sensitivity to them showed to be very much dependent on the precise configuration of the system as a whole.

From the parameter scan, especially of the smaller apoptosis model, and to some extent also by the SASSy and SloppyCell analysis, it became clear that due to the way the two phenotypes had been defined, a sharp edge was formed in the parameter space, separating systems having a healthy phenotype from systems having a diseased phenotype. As the system got closer to this edge that sensitivities of parameters were shifting from the original sensitivity.

Instead of asking under what conditions each genetic variation would be the final driver pushing the system over to a disease phenotype, the question was somewhat reverted; Can

different starting positions in the configuration space be linked to the risk of acquiring a disease phenotype?

Since the two classes of behaviour shared similarities with the mathematical concept of a separatrix in terms of separating the different types of behaviour in parameter space, the term was used throughout this thesis. It is important though to keep in mind that this is a separatrix between sets of model states which can be interpreted as distinct biological phenotypes and not between mathematical basins of attraction.

Any attempt to formulate a biological process, because of its complexity, inevitably will depend on a large number of assumptions about the process. To minimise artefacts resulting from potentially erroneous assumptions, the method developed to link the risk of acquiring a disease phenotype to the distance to the separatrix was kept as simple as possible. Instead of focusing on various aspects of the separatrix, the mean distance was chosen as a single measurement. Furthermore, the boundaries of the state space, corresponding to all possible initial conditions of the model were kept constant when small initial perturbations were introduced.

By applying this new method to the two apoptosis models it could be shown that various parameters had different sensitivities with regards to the distance to the separatrix. That is to say, the change in distance to the separatrix, which a perturbation of similar size had varied between parameters.

A comparison of the outputs of the two models and a closer examination of a subset of the larger model revealed some limitations of using the mean distance in separatrix space to model risk. As the dimensions of the separatrix space increase, the effect any one perturbation will have on the distance to any point on the separatrix will decrease. As carcinogenesis is characterised by a general up-regulation of transcription this poses a problem since one would want to take into account this general up-regulation when modelling the transition from a healthy state to a diseased state. However, as the parameter scans in Chapter 3 showed, there is a wide range of sensitivities with regards to individual components of the network. Thus, at least in the case of the larger apoptosis model studied in this work, there could most likely be a reduction of the total set of variables without sacrificing much of the potential explanatory value of the modelling.

A further limitation of the method was seen when considering large perturbations from the original position in parameter space or when the initial position was relatively close to the separatrix in any one dimension. In these cases, the assumption was no longer valid, that the measured points on the separatrix surface were equally good estimates of the old and the new potential to cross the separatrix and change phenotype. This poses a great limitation on the method, in terms of the range of scenarios which can be studied.

However, the goal of this thesis was never to study the effect of large perturbations on the risk of developing disease. In the cases where the sensitivity of the system is dominated by a few parameters, or where large perturbations are of interest, more standard tools of sensitivity are better suited. The problem we set out to study was how small perturbations affected the risk of phenotype change. If the problem is restricted in this way, the parameters which are highly sensitive could arguably be excluded from the analysis. Likewise the effect of large perturbations are excluded from the study. These perturbations tend to already be well studied as they are easier to model and measure experimentally.

To assess how strongly the distance to the separatrix correlates with the risk of developing a diseased phenotype, a theoretical dataset was created. Genotypes with known effects on the expression levels of the proteins in the two models were simulated. By simulating individuals with these genotypes and having them be exposed to somatic mutations which further perturbed the expression values, before using these expression values in the apoptosis models to calculate the time to onset of apoptosis, the effect each genotype carried on the risk of developing a a cancer phenotype could be established.

The total risk score of the simulated individuals proved to be strongly correlated both with the time to onset of apoptosis and the distance to the separatrix. By introducing noise in the data, it could be shown that only a part of the measured risk score had to be associated with altered expression of the proteins in the model in order to be able to see this correlation.

6.4 Application to Data

When applying the methods developed on experimental data for breast and prostate cancer the results were not as clear as with the simulated data. Neither breast cancer nor prostate cancer associated SNPs could be correlated with an increased distance to the separatrix in the smaller apoptosis model. Breast cancer associated SNPs did have a statistically significant correlation with time to apoptosis in breast tissue, but not on lymphoblastoid cell lines. Since breast cancer associated SNPs were used in the analysis the results using breast tissue is arguably more relevant than the results derived from the lymphoblastoid data. Consequently, the positive results from the former analysis should weigh higher than the negative results from the latter. As was also pointed out in the discussion of Chapter 5, the negative correlation using immune derived cells could be of interest on their own, as breast cancer associated cells have already been predicted to affect the expression of genes with known roles in immune cells [116]. Rendering the cells more sensitive to apoptotic signalling could give the breast cancer cells an increased

chance of evading the immune response. The prostate cancer associated SNPs did not have a significant correlation with time to apoptosis. However, the trend was in the expected direction, with an increased time to apoptosis as the risk score ratio increased.

Using all of the SNPs associated with a cancer type is likely to introduce a lot of noise in the analysis, since only a small fraction of the SNPs are likely to be affecting the expression of the proteins in the apoptosis model. When limiting the analysis to a smaller set of SNPs, there was a significant correlation between the more restricted risk score and the time to apoptosis, as calculated by the model for the breast cancer associated SNPs. However there was still no significant correlation between the risk score and the distance to the separatrix for either cancer type. Close examination of both datasets revealed an enrichment of SNPs with individual effects promoting carcinogenesis among the chosen subset of SNPs compared to the entire set of SNPs. Especially the risk score of the 12 breast cancer associated SNPs had one of the strongest positive correlations with time to apoptosis possible from subsampling 12 SNPs from the original dataset. The correlation between time to apoptosis and risk score for the 7 prostate cancer associated SNPs was also among the most positive ones possible, although the pattern was not as clear. Also the correlation between the risk scores of both cancer types and distance to separatrix was clearly more negative than the average correlation possible from the total sets.

The sample size was smaller for prostate cancer than for breast cancer, with only 50 samples compared to 86 breast tissue samples. This could explain why the breast cancer data set showed stronger correlations than the prostate cancer dataset. Another reason could be that in one dataset a larger fraction of the SNPs were actually associated with altered expression of the proteins in the model than the other dataset. When sample sizes from the theoretical data similar to those available for the two cancer types were used, it was estimated that about 50% of the Risk Score Ratio (RSR) would have to be linked to the model in order to confidently be able to interpret a lack of a correlation in the analysis as an absence of actually correlation between the genotype and the distance to the separatrix. However, the estimation is very conservative as it assumes a perfect correlation between the genotype and the transcription values and a perfect correlation between the transcription values and the translation of the proteins in the model. The subsampling of the SNPs was done based on literature research and the evidence for an association was not always very strong and the effect size of the SNPs was rarely available. It is quite possible that some of the chosen SNPs do not affect the expression values of the proteins in question and that some of the ones that do have an effect, do so in a weak manner.

Overall the trend in the real data is fairly consistent with the simulated results and supports the use of the method. Although, to get stronger results, more care would have to be take to curate the SNPs used or increase the sample size.

6.5 Future Work

In this thesis, the study was limited to two apoptosis models. As was seen in chapter 1 this pathway constitutes a very small, although important, part of the carcinogenesis process. There are many more pathways to study, possibly with better links between genotype and expression values of the proteins involved. One way of increasing the power of the modelling would be to consider models which more accurately model the transcription and translation machinery, or adding a layer between the model and the experimental data to better model these processes.

If the problems arising with higher dimensional separatrix surfaces can be addressed adequately, there is also the possibility of linking several models of different pathways in order to study how they affect each other. For example a cell cycle model could be coupled with an apoptosis model so that the time to apoptosis was more dynamically constrained by a cell cycle which in turn itself was affected by perturbations of proteins involved. Ultimately one would want to link models to all of the Hallmarks of cancer. However, while there are strong links between many of the hallmarks, this is not necessarily the case for all of them. Consequently, one would not need to incorporate all of the hallmarks in one single separatrix surface.

These are just some of the possible areas for future work. We have established an initial framework which can be used to examine the effect of SNPs that alter gene expression on the risk of developing a cancer phenotype. Dynamical mathematical models have previously been used to study the progression from healthy phenotype to cancer. However, to our knowledge, this is the first time they have been used in combination with SNP data in an attempt to study the mechanistics behind the risk of developing cancer. Even though the results presented here were quite modest with regards to being able to apply a theoretical framework to experimental data and gain understanding of the biological process, it is my hope that it will serve as an inspiration and point towards the possibilities in combining mathematical modelling and experimental data to explore the genotype-phenotype landscape with regards to risk of developing a disease. I have no doubt that the continuously more sophisticated mathematical models being developed paired with the accumulation of more genotype and transcription data will strongly improve the power

of the tools developed in this thesis and any other methods developed in the future by people posing similar questions.

Appendix A

Breast Cancer SNPs Used During Data Mining

Table A.1: Breast cancer associated SNPs.

rs ID	Chromosome	Risk Allele	OR	logOR
rs616488	1	A	0.94	0.061875
rs11552449	1	T	1.07	0.067659
rs11249433	1	G	1.16	0.148420
rs4849887	2	C	0.91	0.094311
rs2016394	2	G	0.95	0.051293
rs1550623	2	A	0.94	0.061875
rs13387042	2	A	1.19	0.173953
rs16857609	2	T	1.08	0.076961
rs6762644	3	G	1.07	0.067659
rs4973768	3	T	1.15	0.139762
rs12493607	3	C	1.06	0.058269
rs6788895	3	G	1.22	0.198851
rs9790517	4	T	1.05	0.048790
rs6828523	4	C	0.90	0.105361
rs10069690	5	T	1.18	0.165514
rs1092913	5	A	1.45	0.371564
rs4415084	5	T	1.17	0.157004
rs7716600	5	A	1.24	0.215111
rs16886165	5	G	1.23	0.207014
rs889312	5	C	1.17	0.157004
rs10472076	5	C	1.05	0.048790
rs1353747	5	T	0.92	0.083382
rs1432679	5	C	1.07	0.067659
rs11242675	6	T	0.94	0.061875
rs204247	6	G	1.05	0.048790

Table A.1: (continued)

rs17530068	6	C	1.16	0.148420
rs2180341	6	G	1.41	0.343590
rs9485372	6	G	1.11	0.104360
rs3757318	6	A	1.30	0.262364
rs3734805	6	C	1.19	0.173953
rs2046210	6	A	1.29	0.254642
rs9383938	6	T	1.28	0.246860
rs9383951	6	G	1.14	0.131028
rs2048672	7	C	1.11	0.104360
rs720475	7	G	0.94	0.061875
rs9693444	8	A	1.07	0.067659
rs6472903	8	T	0.91	0.094311
rs2943559	8	G	1.13	0.122218
rs13281615	8	G	1.08	0.076961
rs1562430	8	T	1.16	0.148420
rs11780156	8	T	1.07	0.067659
rs1011970	9	T	1.09	0.086178
rs10759243	9	A	1.06	0.058269
rs865686	9	T	1.12	0.113329
rs2380205	10	C	1.06	0.058269
rs7072776	10	A	1.07	0.067659
rs11814448	10	C	1.26	0.231112
rs10822013	10	T	1.12	0.113329
rs10995190	10	G	1.16	0.148420
rs704010	10	T	1.07	0.067659
rs7904519	10	G	1.06	0.058269
rs11199914	10	C	0.95	0.051293
rs3750817	10	T	1.22	0.198851
rs2981579	10	A	1.30	0.262364
rs2981582	10	A	1.26	0.231112
rs10510102	10	C	1.12	0.113329
rs3817198	11	C	1.07	0.067659
rs909116	11	T	1.17	0.157004
rs3903072	11	G	0.95	0.051293
rs614367	11	T	1.15	0.139762
rs11820646	11	C	0.95	0.051293
rs7107217	11	C	1.08	0.076961
rs12422552	12	C	1.05	0.048790
rs17356907	12	A	0.91	0.094311
rs11571833	13	T	1.26	0.231112
rs2236007	14	G	0.93	0.072571

Table A.1: (continued)

rs2588809	14	T	1.08	0.076961
rs999737	14	C	1.06	0.058269
rs4322600	14	G	1.18	0.165514
rs941764	14	G	1.06	0.058269
rs3803662	16	A	1.23	0.207014
rs4784227	16	T	1.24	0.215111
rs3112612	16	A	1.15	0.139762
rs17817449	16	T	0.93	0.072571
rs13329835	16	G	1.08	0.076961
rs527616	18	G	0.95	0.051293
rs1436904	18	T	0.96	0.040822
rs8170	19	A	1.26	0.231112
rs8100241	19	G	1.14	0.131028
rs4808801	19	A	0.93	0.072571
rs3760982	19	A	1.06	0.058269
rs10411161	19	T	1.42	0.350657
rs2284378	20	T	1.16	0.148420
rs132390	22	C	1.12	0.113329
rs6001930	22	C	1.12	0.113329

Table A.2: Gene names mapping to proteins in the boolean cancer development model by Fumiã & Martins [53].

AKT1	CHEK2	MDM2	SLC2A1
AMP	COX4I2	MTOR	SMAD1
ATM	DVL1	MXI1	SMAD2
ATP	E2F1	MYC	SMAD3
ATR	E2F2	NF1	SMAD4
APAF1	E2F3	NFKB1	SMAD5
APC	E2F4	NFKB2	SMAD6
ARAF	E2F5	PDK1	SMAD7
BAD	E2F6	PIK3CD	SMAD9
BAK1	E2F7	PIP3	SNAI1
BAX	E2F8	PRKAA1	SNAI2
BCAT1	EEF2	PRKAA2	SSSCA1
BCAT2	EEF2K	PRKAB1	TCF7
BCL2	FADD	PRKAB2	TCF7L1
BCL2L	FOS	PRKAG1	TCF7L2
BRAF	FOXO1	PRKAG2	TERC
CASP8	FOXO3	PRKAG3	TERT
CASP9	FOXO4	PKRCA	TGFB1
CCNA1	FOXO6	PRKCB	TNF
CCNA2	GLI1	PTEN	TP53
CCNB1	GSK	RAF1	TSC1
CCNB2	GSR	RAG1	TSC2
CCND1	GSS	RAG2	UBE2C
CCND2	HIF1A	RB1	VEGFA
CCND3	IKBKB	RHEB	VEGFB
CCNE1	JUN	RPS6KA1	VEGFC
CCNE2	KRAS	RPS6KA2	VHL
CDC20	LDHA	RPS6KA3	WNT1
CDH1	MAP3K7	RPS6KB1	ZMIZ1
CDKN1A	MAPK1	RPS6KB2	
CDKN2B	MAPK8	RTK	
CHEK1	MAX	SF3B6	

Table A.3: Tabular form of interactions between genes targeted by eQTLs and proteins in the Boolean cancer model as depicted by Figure 2.3.

eQTL target gene	intermediate proteins	model protein
EIF2S2	RPS3, RPS5, RPS6, RPS7, RPS8, RPS9, RPS10, RPS13, RPS14, RPS15A, RPS16, RPS19, RPS20, RPS23, RPS24, RPS29, FAU	EEF2
ELL	CDK7, MNAT1, CCNH	CCNB1, CCND1, CCNE1, CCNE2, CDKN1A
ELL	GTF2F2, GTF2F1, POLR2B, POLR2C, POLR2D, POLR2E, POLR2F, POLR2G, POLR2I, POLR2L	SF3B6
ELL	TCEB1, TCEB2	HIF1A, VHL
PLAUR	PLG, SERPINE1	TGFB1
TNNT3	ACTN2	TGFB1
CHMP4B, TGFR2	UBA52	EEF2
CHMP4B, TGFR2	UBA52, UBB, UBC	SMAD4, SMAD7, TGFB1 CDC20, CDKN1A, HIF1A, UBE2C, VHL
TGFBR2	CGN, PARD6A, PARD3, UCHL5, ARHGEF18, PPP1CC, TGFBR1, PPP1CB, NED4L, USP15, SMURF2	TGFB1
TGFBR2	SMURF2 NED4L	SMAD4
TGFBR2	UCHL5, SMURF2 PPP1CC, TGFBR1, PPP1CB, NED4L, USP15	SMAD7
TGFBR2	SMURF2	SMAD9

Table A.4: Tabular form of interactions betwee breast cancer associated SNPs and proteins in the Boolean cancer model as depicted in Figure 2.5.

SNP target gene	intermediate proteins	model protein
CCND1	UBC, UBA52, UBB	SMAD4, SMAD7, VHL, CDC20, HIF1A, UBE2C, TGFB1
CCND1	CDK4	CCND1, E2F5, E2F1, E2F4, E2F2, RB1, E2F3, ATM
CCND1	CDKN1B, TFDP1	E2F3, RB1, E2F2, E2F1, E2F5, E2F4
CCND1	RBL1	E2F4
CCND1	RBL2	E2F4, E2F5
CCND1	-	CDKN1A, E2F1, E2F2, E2F3, E2F4, E2F5, RB1
EIF2S2	RPS9, RPS5, RPS16, RPS15A, RPS29, RPS20, FAU, RPS3, RPS7, RPS19, RSP13, RPS6, RPS8, RPS10, RPS23, RPS24, RPS14	EEF2
ESR1	-	TNF
MAP3k	CHUK	NFKB1, NFKB2, MAP3K7, IKBKB
MAP3K1	-	IKBKB
RAD23B	RPA3	ATM
RAD23B	CCNH, MNAT1, CDK7	CCND1, CDKN1A, CCNE1, CCNB1, CCNE2
TERT	DKC1	TERT

Appendix B

Smaller Apoptosis Model

$$\frac{d[C8]}{dt} = -k_2[C3^*][C8] - k_9[C8] + k_{-9} \tag{B.1}$$

$$\frac{d[C8^*]}{dt} = k_2[C3^*][C8] - k_5[C8^*] - k_{11}[C8^*][BAR] + k_{-11}[C8^* \sim BAR] \tag{B.2}$$

$$\frac{d[C3]}{dt} = -k_1[C8^*][C3] - k_{10}[C3] + k_{-10} \tag{B.3}$$

$$\frac{d[C3^*]}{dt} = k_1[C8^*][C3] - k_3[C3^*][IAP] + k_{-3}[C3^* \sim IAP] - k_6[C3^*] \tag{B.4}$$

$$\frac{d[IAP]}{dt} = -k_3[C3^*][IAP] + k_{-3}[C3^* \sim IAP] - k_4[C3^*][IAP] - k_8[IAP] + k_{-8} \tag{B.5}$$

$$\frac{d[C3^* \sim IAP]}{dt} = k_3[C3^*][IAP] - k_{-3}[C3^* \sim IAP] - k_7[C3^* \sim IAP] \tag{B.6}$$

$$\frac{d[BAR]}{dt} = -k_{11}[C8^*][BAR] + k_{-11}[C8^* \sim BAR] - k_{12}[BAR] + k_{-12} \tag{B.7}$$

$$\frac{d[C8^* \sim BAR]}{dt} = k_{11}[C8^*][BAR] - k_{-11}[C8^* \sim BAR] - k_{13}[C8^* \sim BAR] \tag{B.8}$$

Table B.1: Standard parameter settings of the smaller apoptosis model.

Parameter name	Value	Unit
k_1	5.8E-5	$cell \cdot mol^{-1} \cdot min^{-1}$
k_2	1.0E-5	$cell \cdot mol^{-1} \cdot min^{-1}$
k_3	5.0E-4	$cell \cdot mol^{-1} \cdot min^{-1}$
k_4	3.0E-4	$cell \cdot mol^{-1} \cdot min^{-1}$
k_5	5.8E-3	min^{-1}
k_6	5.8E-3	min^{-1}
k_7	1.73E-2	min^{-1}
k_8	1.16E-2	min^{-1}
k_9	3.9E-3	min^{-1}
k_{10}	3.9E-3	min^{-1}
k_{11}	5.0E-4	$cell \cdot mol^{-1} \cdot min^{-1}$
k_{12}	1.0E-3	min^{-1}
k_{13}	1.16E-2	min^{-1}
k_{-3}	0.21	min^{-1}
k_{-8}	464	$cell \cdot mol^{-1} \cdot min^{-1}$
k_{-9}	507	$cell \cdot mol^{-1} \cdot min^{-1}$
k_{-10}	81.9	$cell \cdot mol^{-1} \cdot min^{-1}$
k_{-11}	0.21	min^{-1}
k_{-12}	40	$cell \cdot mol^{-1} \cdot min^{-1}$

Table B.2: Standard variable settings of the smaller apoptosis model. All values are in molecules/cell.

Variable name	Value
C8	130000.0
C8*	1000.0
C3	21000.0
C3*	0.0
BAR	40000.0
BAR-C8*	0.0
IAP	40000.0
IAP-C3*	0.0

Appendix C

Larger Apoptosis Model

$$\frac{d[com0]}{dt} = -k_{19}[TNF][com0] + k_{43} * incom0 * (1 - actD) - k_{43}[com0] \tag{C.1}$$

$$\frac{d[com1]}{dt} = k_{19}[TNF][com0] - k_{20}[com1][FADD] - k_{44}[com1] \tag{C.2}$$

$$\frac{d[proMKK7]}{dt} = -k_{23}[com1][proMKK7] + k_{27}[MKK7][phos.] \tag{C.3}$$

$$\frac{d[MKK7]}{dt} = k_{23}[com1][proMKK7] - k_{27}[MKK7][phos.] \tag{C.4}$$

$$\frac{d[JNK]}{dt} = -k_{24}[MKK7][JNK] + k_{34}[pJNK][MKP] + k_{33}[ROS] \tag{C.5}$$

$$\frac{d[pJNK]}{dt} = k_{24}[MKK7][JNK] - k_{34}[pJNK][MKP] \tag{C.6}$$

$$\frac{d[prophos.]}{dt} = -k_{25}[prophos.][pJNK] + k_{26}[phos.] \tag{C.7}$$

$$\frac{d[phos.]}{dt} = k_{25}[prophos.][pJNK] - k_{26}[phos.] \tag{C.8}$$

$$\frac{d[FADD]}{dt} = -k_{20}[com1][FADD] - k_{47}[Fas][proD][FADD] \tag{C.9}$$

$$\frac{d[proD]}{dt} = -k_{47}[Fas][proD][FADD] \tag{C.10}$$

$$\frac{d[C8]}{dt} = -k_{21}[com2][C8] - k_{22}[D][C8] \tag{C.11}$$

$$\frac{d[C8a]}{dt} = k_{21}[com2][C8] + k_{22}[D][C8] - k_8[C8a] \tag{C.12}$$

$$\frac{d[c2]}{dt} = k_{20}[c1][FADD] - k_{39}[c2][cF] + k_{41}[c2F] - k_{45}[c2] \tag{C.13}$$

$$\frac{d[D]}{dt} = -k_{40}[D][cF] + k_{42}[DcF] - k_{46}[D] + k_{47}[Fas][proD][FADD] \tag{C.14}$$

$$\frac{d[c2F]}{dt} = k_{39}[com2][cF] - k_{41}[c2F] \tag{C.15}$$

$$\frac{d[DcF]}{dt} = k_{40}[D][cF] - k_{42}[DcF] \tag{C.16}$$

$$\frac{d[cF]}{dt} = -k_{38}[cF][iPPP] - k_{39}[c2][cF] - k_{40}[D][cF] + k_{41}[c2F] + k_{42}[DcF] \tag{C.17}$$

$$\frac{d[itch]}{dt} = -k_{35} * CHX[itch][pJNK] \tag{C.18}$$

$$\frac{d[itchP]}{dt} = k_{35} * CHX[itch][pJNK] - k_{36}[itchP][pJNK] \tag{C.19}$$

$$\frac{d[itchPP]}{dt} = k_{36}[itchP][pJNK] - k_{37}[itchPP][pJNK] \tag{C.20}$$

$$\frac{d[itchPPP]}{dt} = k_{37}[itchPP][pJNK] \tag{C.21}$$

$$\frac{d[IKKn]}{dt} = k_{prod} * (1 - actD) - k_{deg}[IKKn] - T_R * k_1[IKKn] \tag{C.22}$$

$$\frac{d[A20_t]}{dt} = c_2 * (1 - actD) + c_1[NF\kappa B_n] * (1 - actD) - c_3[A20_t] \tag{C.23}$$

$$\frac{d[I\kappa B\alpha_t]}{dt} = c_{2\alpha} * (1 - actD) + c_{1\alpha}[NF\kappa B_n] * (1 - actD) - c_{3\alpha}[I\kappa B\alpha_t] \tag{C.24}$$

$$\frac{d[PmRNA]}{dt} = c_{1c}[NF\kappa B_n] * (1 - actD) - c_{2c}[PmRNA] \tag{C.25}$$

$$\frac{d[P]}{dt} = k_{28}[PmRNA] - k_{30}[P][ROS] \tag{C.26}$$

$$\frac{d[MKP]}{dt} = -k_{31}[ROS][MKP] + k_{32}[MKPox] \tag{C.27}$$

$$\frac{d[MKPox]}{dt} = k_{31}[ROS][MKP] - k_{32}[MKPox] \tag{C.28}$$

$$\frac{d[ROS]}{dt} = [ROSfree] - k_{30}[P][ROS] - k_{31}[ROS][MKP] \tag{C.29}$$

$$\frac{d[JNK]}{dt} = -k_1[TNF][JNK] \tag{C.30}$$

$$\frac{d[pJNK]}{dt} = k_1[TNF][JNK] - k_2[pJNK] \tag{C.31}$$

$$\frac{d[Bim]}{dt} = -k_3[pJNK][Bim] \tag{C.32}$$

$$\frac{d[pBim]}{dt} = k_3[pJNK][Bim] - k_4[pBim] - k_5[pBim][Bcl2] \tag{C.33}$$

$$\frac{d[C8]}{dt} = -k_{21}[com2][C8] - k_{22}[D][C8] \tag{C.34}$$

$$\frac{d[C8^*]}{dt} = k_{21}[com2][C8] + k_{22}[D][C8] - k_8[C8^*] \tag{C.35}$$

$$\frac{d[C3]}{dt} = -k_{14}[C3][C8^*] - k_{15}[C3][CytC_{free}] - k_{16}[C3][C3^*] \tag{C.36}$$

$$\frac{d[C3^*]}{dt} = k_{14}[C3][C8^*] + k_{15}[C3][CytC_{free}] + k_{16}[C3][C3^*] - k_{17}[C3^*] - k_{18}[C3^*][XIAP] \tag{C.37}$$

$$\frac{d[Bid]}{dt} = -k_9[C8^*][Bid] \tag{C.38}$$

$$\frac{d[tBid]}{dt} = k_9[C8^*][Bid] - k_{10}[tBid] - k_{11}[tBid][Bcl2] \tag{C.39}$$

$$\frac{d[BaxBak]}{dt} = -k_6[pBim][BaxBak] - k_{12}[tBid][BaxBak] + k_{13}[BaxBak^*] \tag{C.40}$$

$$\frac{d[BaxBak^*]}{dt} = k_6[pBim][BaxBak] + k_{12}[tBid][BaxBak] - k_{13}[BaxBak^*] \tag{C.41}$$

$$\frac{d[XIAP]}{dt} = -k_{18}[C3^*][XIAP] \tag{C.42}$$

$$\frac{d[Bcl2]}{dt} = -k_5[pBim][Bcl2] - k_{11}[tBid][Bcl2] \tag{C.43}$$

$$Fas(t) = H(t-a) * 100 \tag{C.44}$$

$$CytC_{free}(t) = H(BaxBak^*(t) - 90) * 100 \tag{C.45}$$

$$\text{Where: } H(n) = \begin{cases} 0, & n < 0 \\ 1, & n \geqq 0 \end{cases} \tag{C.46}$$

$$ROS_{free}(t) = \frac{1}{0.03 * 2\pi} e^{\frac{1}{2}(\frac{t-4}{0.03})^2} * 100 * (1 - BHA) \tag{C.47}$$

Table C.1: Initial variable settings of the larger apoptosis model. All values are in arbitrary units (AU)

Variable name	Value	complex0	100
TNF	100	complex1	0
JNK	100	complex2	0
pJNK	0	cFLIP	100
Bim	100	c2FLIP	0
pBim	0	P	0
C8	100	ROS	0
C8*	0	MKP	50
C3	100	MKPox	0
C3*	0	proMKK7	100
Bid	100	MKK7	0
tBid	0	prophosphatase	100
BaxBak	100	phosphatase	0
BaxBak*	0	itch	100
XIAP	80	itchP	0
Bcl2	100	itchPP	0
FADD	200	itchPPP	0
proDISC	100	DISC	0
DcFLIP	0	cgent	0.0
IKKn	0.2	A20t	0.0
IkBat	0.0	PmRNA	0.0
IKKa	0.0	IKKi	0.0
IKKaIkBa	0.0	IKKaIkBaNFkB	0.0
NFkB	0.00033705498019754324	NFkBn	0.002203216237184229
A20	0.004590033827467142	IkBa	0.0019900111439232052
IkBan	0.002294747386858693	IkBaNFkB	0.05890045169239104
IkBanNFkBn	0.00008426374504938584		

Table C.2: Standard parameter settings of the larger apoptosis model.

Parameter name	Value	Unit	Parameter name	Value	Unit
k_3	0.04	$AU^{-1}h^{-1}$	k_{45}	0.05	h^{-1}
k_4	0.001	h^{-1}	k_{46}	0.05	h^{-1}
k_5	1.0	$AU^{-1}h^{-1}$	k_{47}	0.005	$AU^{-2}h^{-1}$
k_6	0.005	$AU^{-1}h^{-1}$	t_1	360.0	h^{-1}
k_8	0.01	h^{-1}	t_2	360.0	h^{-1}
k_9	0.002	$AU^{-1}h^{-1}$	c_{1a}	0.0018	h^{-1}
k_{10}	0.001	h^{-1}	c_{2a}	0.0	$AU^{-1}h^{-1}$
k_{11}	1.0	$AU^{-1}h^{-1}$	c_{3a}	1.44	h^{-1}
k_{12}	0.1	$AU^{-1}h^{-1}$	c_{4a}	1800	h^{-1}
k_{13}	0.0001	h^{-1}	c_{5a}	0.36	h^{-1}
k_{14}	0.002	$AU^{-1}h^{-1}$	c_{6a}	0.072	h^{-1}
k_{15}	0.05	$AU^{-1}h^{-1}$	c_1	0.0018	h^{-1}
k_{16}	0.007	$AU^{-1}h^{-1}$	c_2	0.0	$AU^{-1}h^{-1}$
k_{17}	0.01	h^{-1}	c_3	1.44	h^{-1}
k_{18}	0.05	$AU^{-1}h^{-1}$	c_4	1800	h^{-1}
k_{19}	0.05	$AU^{-1}h^{-1}$	c_5	1.08	h^{-1}
k_{20}	0.001	$AU^{-1}h^{-1}$	i_{k1}	9.0	h^{-1}
k_{21}	0.08	$AU^{-1}h^{-1}$	i_{k2}	360	h^{-1}
k_{22}	0.8	$AU^{-1}h^{-1}$	i_{k3}	5.4	h^{-1}
k_{23}	0.05	$AU^{-1}h^{-1}$	k_{prod}	0.09	$AU^{-1}h^{-1}$
k_{24}	0.4	$AU^{-1}h^{-1}$	k_{deg}	0.45	h^{-1}
k_{25}	0.05	$AU^{-1}h^{-1}$			
k_{26}	0.05	h^{-1}	k_v	3.0	-
k_{27}	2.0	$AU^{-1}h^{-1}$	i_1	9.0	h^{-1}
k_{28}	90000	h^{-1}	i_{1a}	3.6	h^{-1}
k_{30}	1.0	$AU^{-1}h^{-1}$	e_{1a}	1.8	h^{-1}
k_{31}	0.1	$AU^{-1}h^{-1}$	e_{2a}	36.0	h^{-1}
k_{32}	0.01	h^{-1}	c_{1c}	0.0018	h^{-1}
k_{33}	2.0	h^{-1}	c_{2c}	0.36	$AU^{-1}h^{-1}$
k_{34}	0.9	$AU^{-1}h^{-1}$	c_{3c}	1.44	h^{-1}
k_{35}	0.015	$AU^{-1}h^{-1}$	$time$	12.0	h
k_{36}	0.025	$AU^{-1}h^{-1}$	$ActD$	1	$AU^{-1}h^{-1}$
k_{37}	0.045	$AU^{-1}h^{-1}$	TNF	100.0	$AU^{-1}h^{-1}$
k_{38}	0.05	$AU^{-1}h^{-1}$	BHA	0	$AU^{-1}h^{-1}$
k_{39}	0.8	$AU^{-1}h^{-1}$	CHX	0	$AU^{-1}h^{-1}$
k_{40}	8.0	$AU^{-1}h^{-1}$	$incomp0$	100	AU
k_{41}	0.008	h^{-1}	a_1	1800	$AU^{-1}h^{-1}$
k_{42}	0.008	h^{-1}	a_2	720.0	$AU^{-1}h^{-1}$
k_{43}	0.001	h^{-1}	a_3	3600.0	$AU^{-1}h^{-1}$
k_{44}	0.05	h^{-1}			

Appendix D

Sensitivity Analysis of the Smaller Apoptosis Model

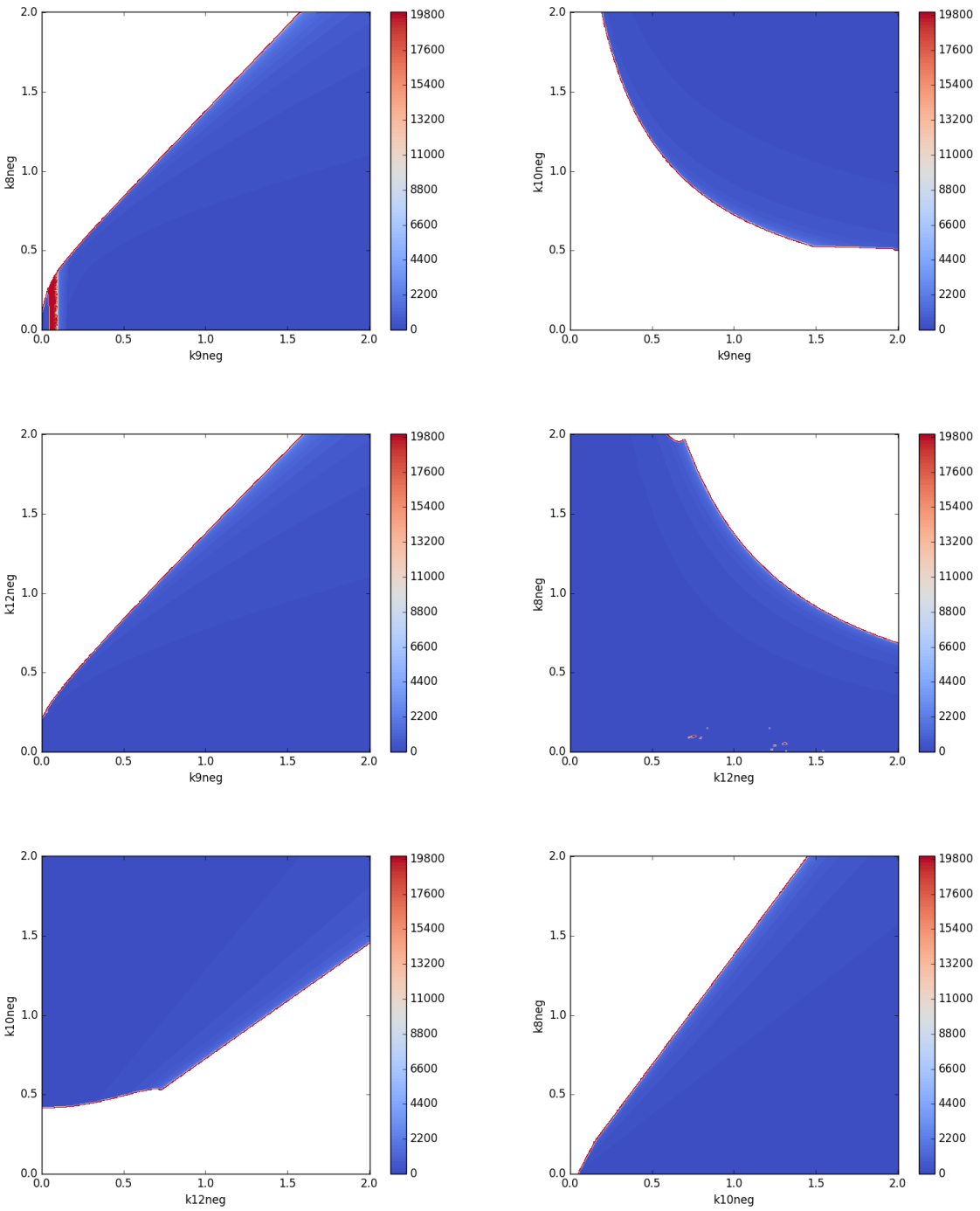


Figure D.1: Time to maximum Caspase 3 signalling when perturbing two parameters between 0 and 2 time the initial value. The value of the parameters is depicted on the respective axis and time time is colour coded from 0 (dark blue) to 20,000 (dark red). Upon initial perturbations, the time to apoptosis does not alter much. However, within a very small window of parameter perturbation, the time changes from very short, to very long. The white indicates where the value exceeded the limit of the scale.

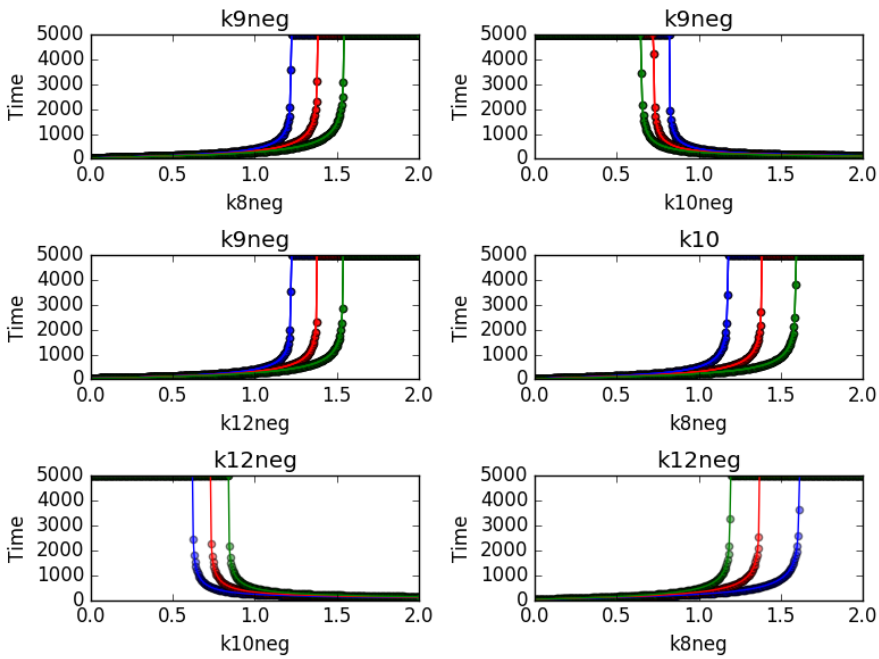
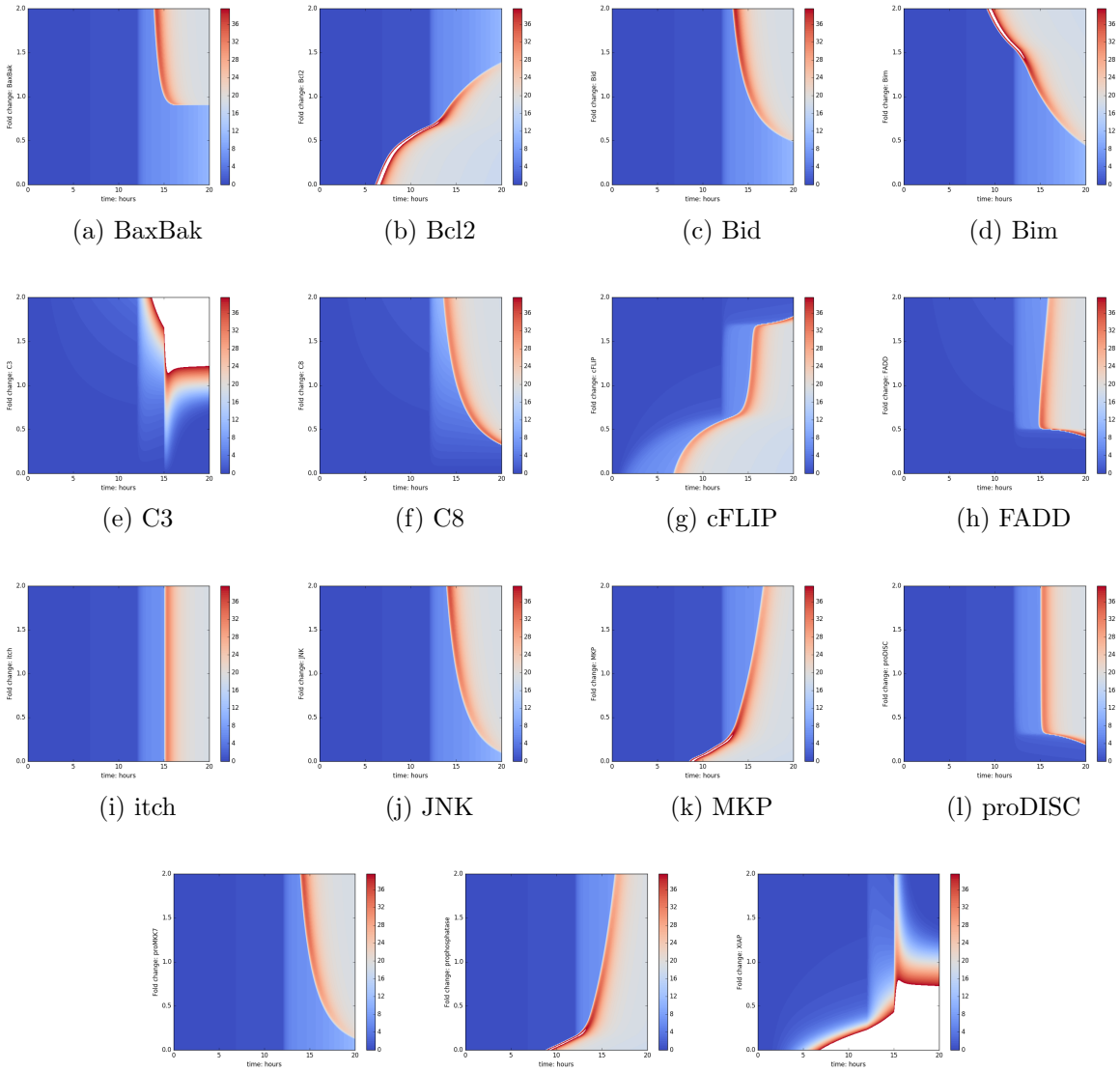


Figure D.2: Time to apoptosis on the y-axis as a function of parameter perturbation. Blue and green lines show an additional 10% perturbation of a second parameter whereas red show the initial value.

Appendix E

Sensitivity Analysis of the Larger Apoptosis Model



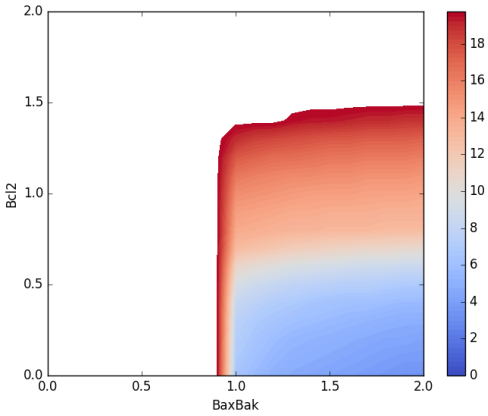
(a) BaxBak (b) Bcl2 (c) Bid (d) Bim

(e) C3 (f) C8 (g) cFLIP (h) FADD

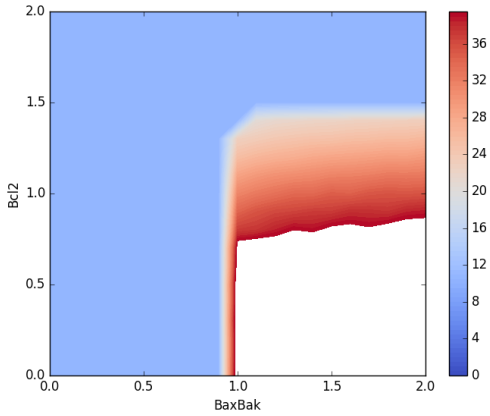
(i) itch (j) JNK (k) MKP (l) proDISC

(m) proMKK7 (n) prophospatase (o) XIAP

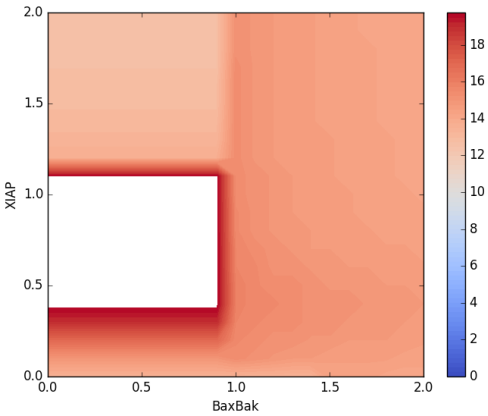
Figure E.1: Concentration of activated Caspase 3 in the larger model over time as one initial concentration if permuted. Colours indicate concentration from 0 (dark blue) to 36 (deep red) and the concentration of the perturbed variable is indicated on the y-axis from 0-2 times normal value. The white indicates where the value exceeded the limit of the scale.



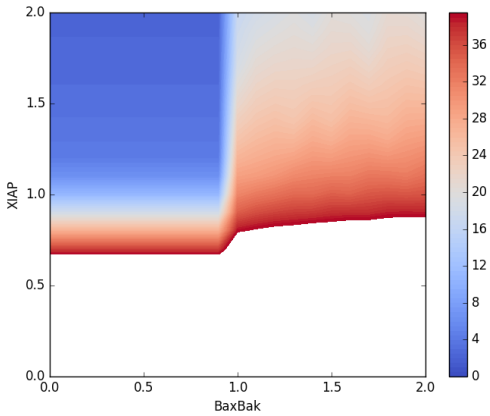
(a) time to max C3*: Bcl2-BaxBak



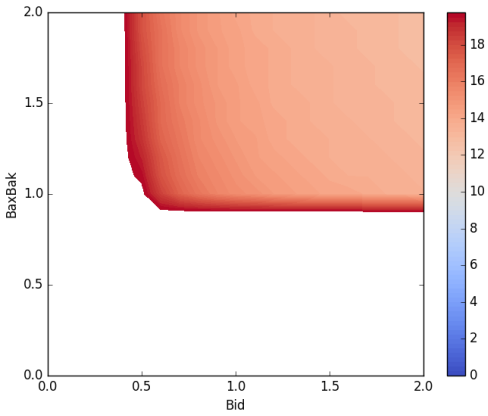
(b) max C3*: Bcl2-BaxBak



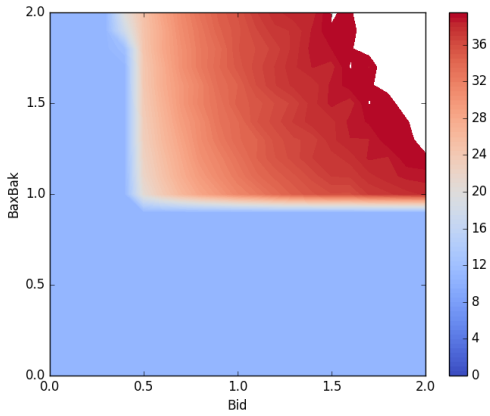
(c) time to max C3*: XIAP-BaxBak



(d) max C3*: XIAP-BaxBak



(e) time to max C3*: BaxBak-Bid



(f) max C3*: BaxBak-Bid

Figure E.2: Left: time to maximum concentration of activated Caspase 3 as two initial concentrations are perturbed. Right: Maximum concentration of activated Caspase 3 within 20 hours as the same initial concentrations are perturbed. The white indicates where the value exceeded the limit of the scale.

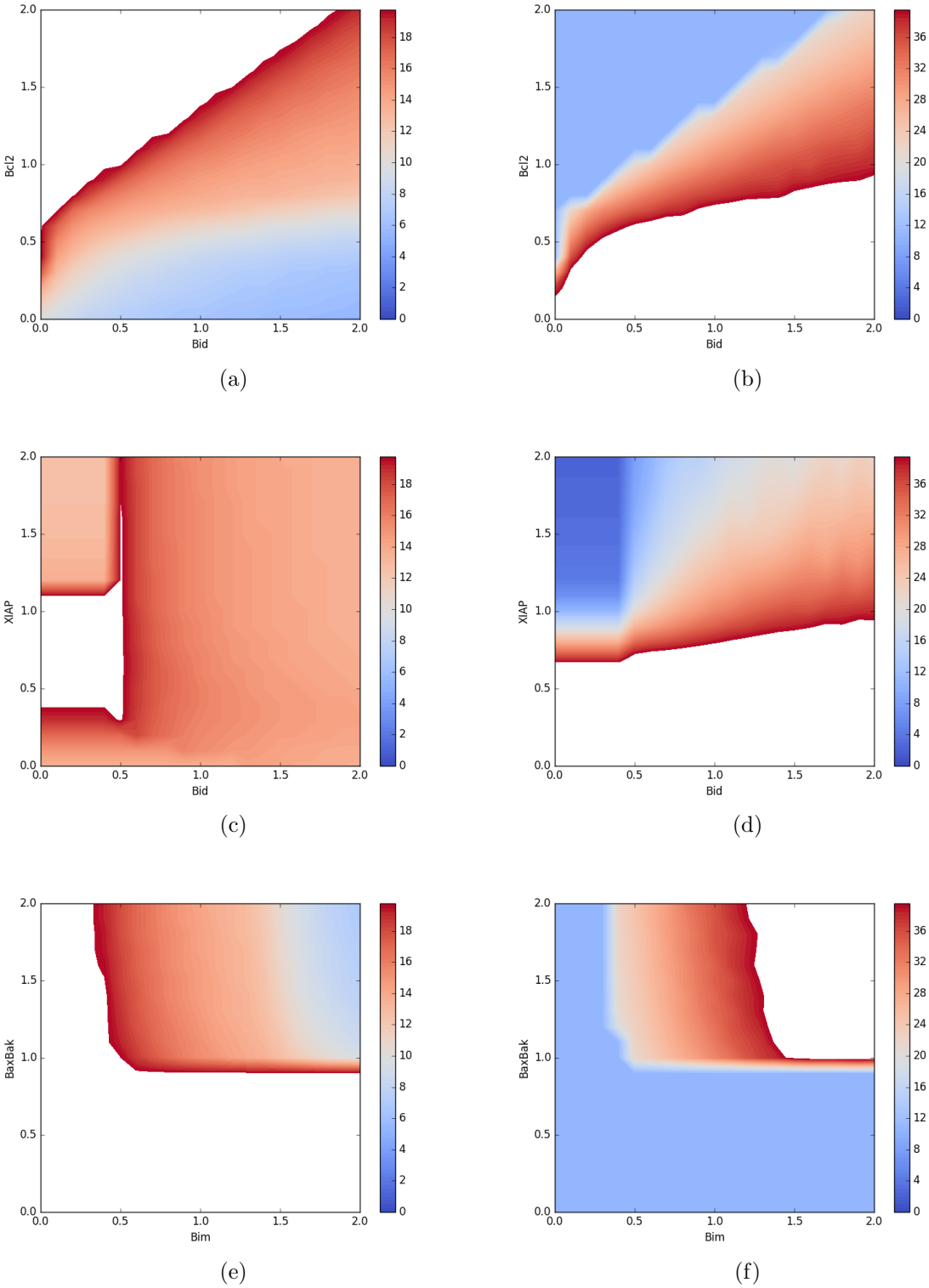


Figure E.3: Left: time to maximum concentration of activated Caspase 3 as two initial concentrations are perturbed. Right: Maximum concentration of activated Caspase 3 within 20 hours as the same initial concentrations are perturbed. The white indicates where the value exceeded the limit of the scale.

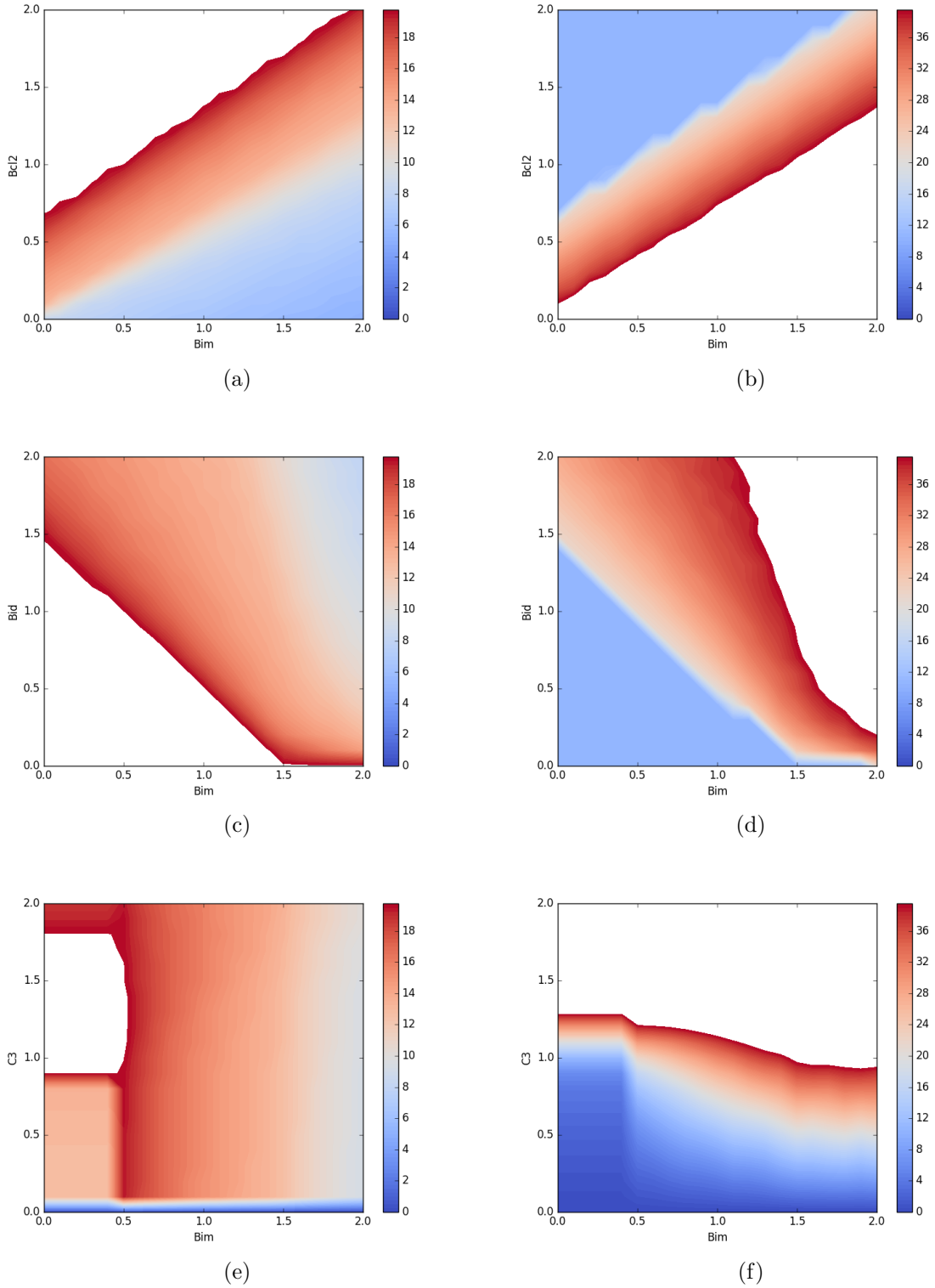


Figure E.4: Left: time to maximum concentration of activated Caspase 3 as two initial concentrations are perturbed. Right: Maximum concentration of activated Caspase 3 within 20 hours as the same initial concentrations are perturbed. The white indicates where the value exceeded the limit of the scale.

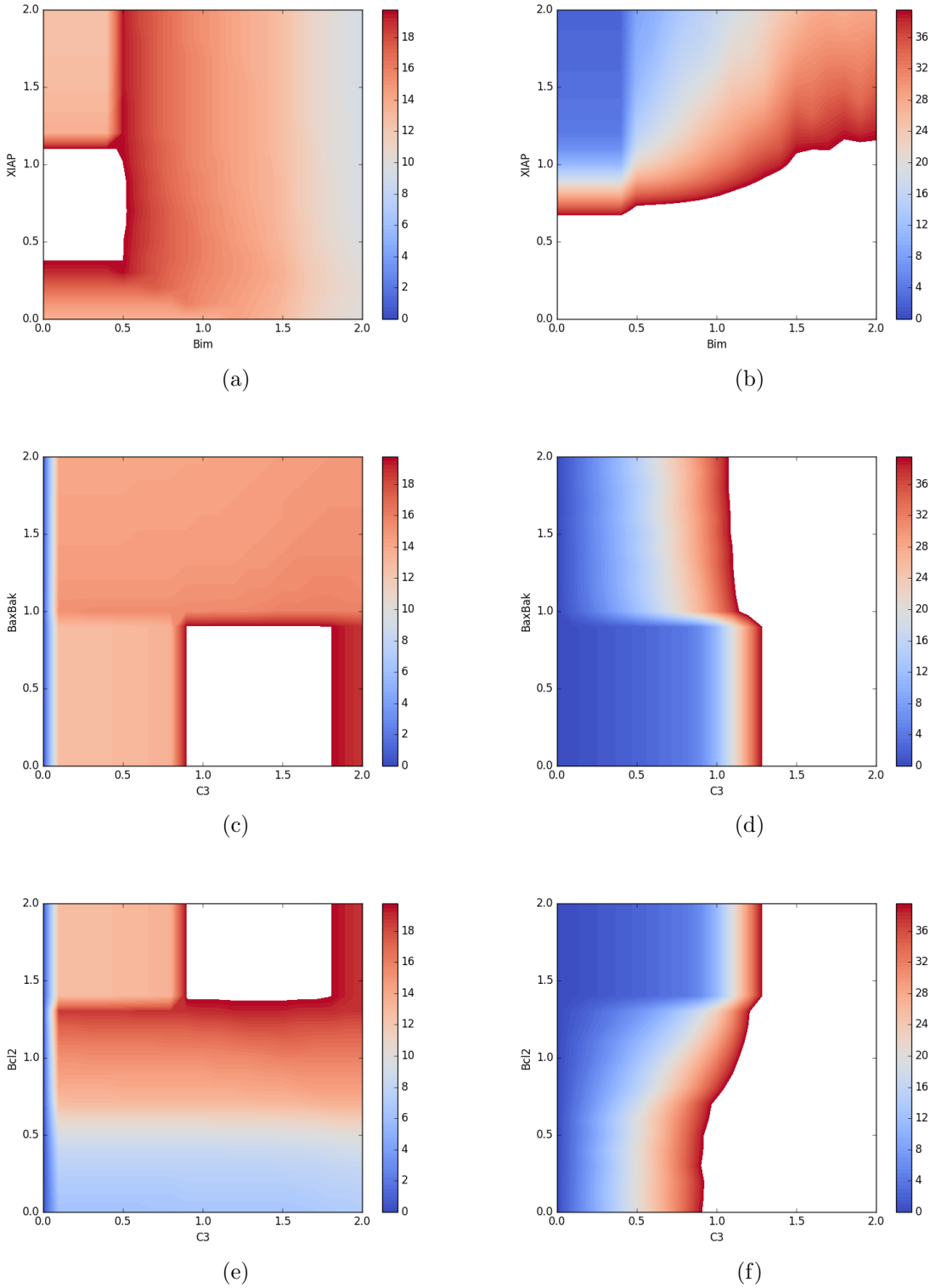


Figure E.5: Left: time to maximum concentration of activated Caspase 3 as two initial concentrations are perturbed. Right: Maximum concentration of activated Caspase 3 within 20 hours as the same initial concentrations are perturbed. The white indicates where the value exceeded the limit of the scale.

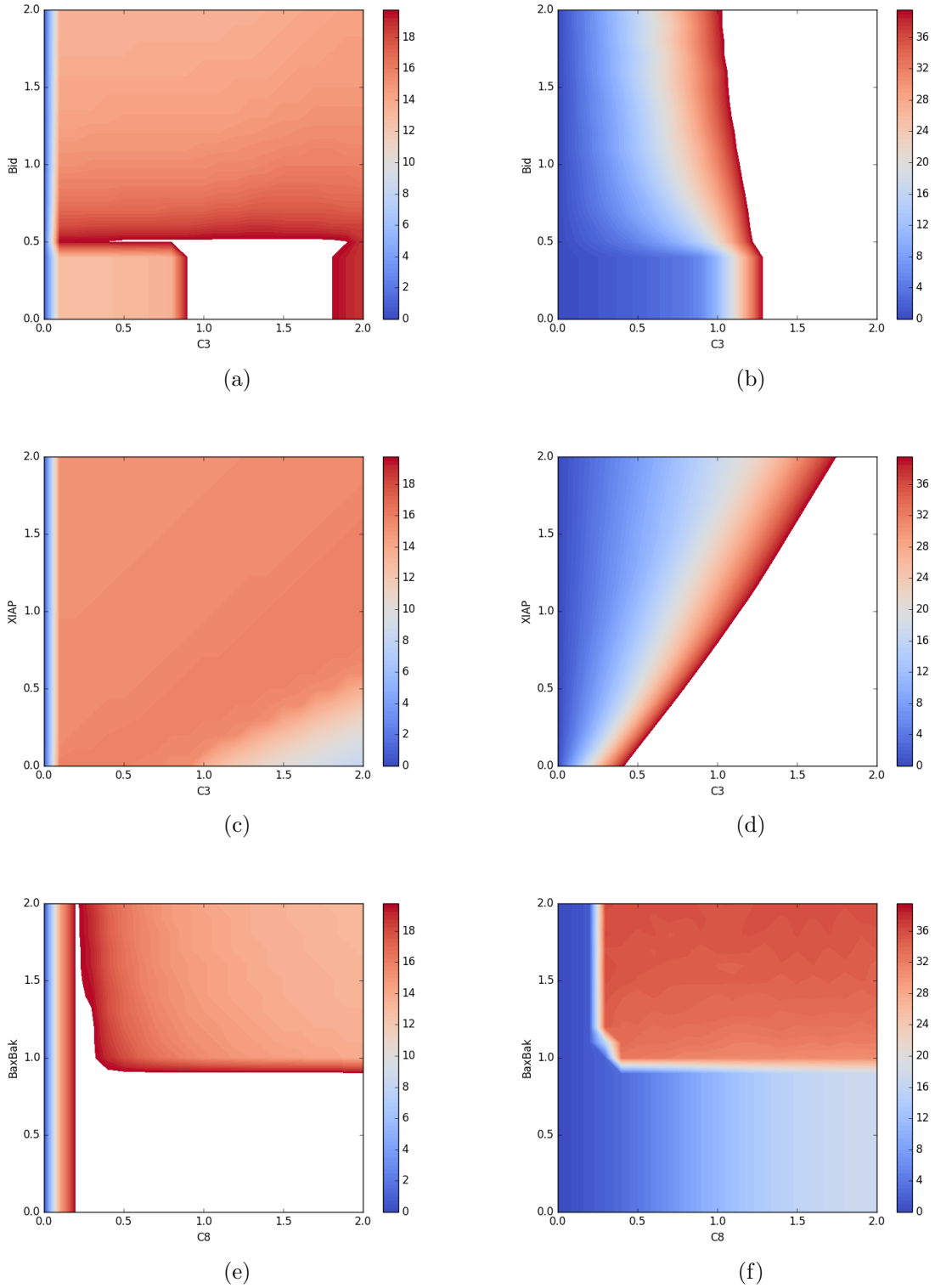


Figure E.6: Left: time to maximum concentration of activated Caspase 3 as two initial concentrations are perturbed. Right: Maximum concentration of activated Caspase 3 within 20 hours as the same initial concentrations are perturbed. The white indicates where the value exceeded the limit of the scale.

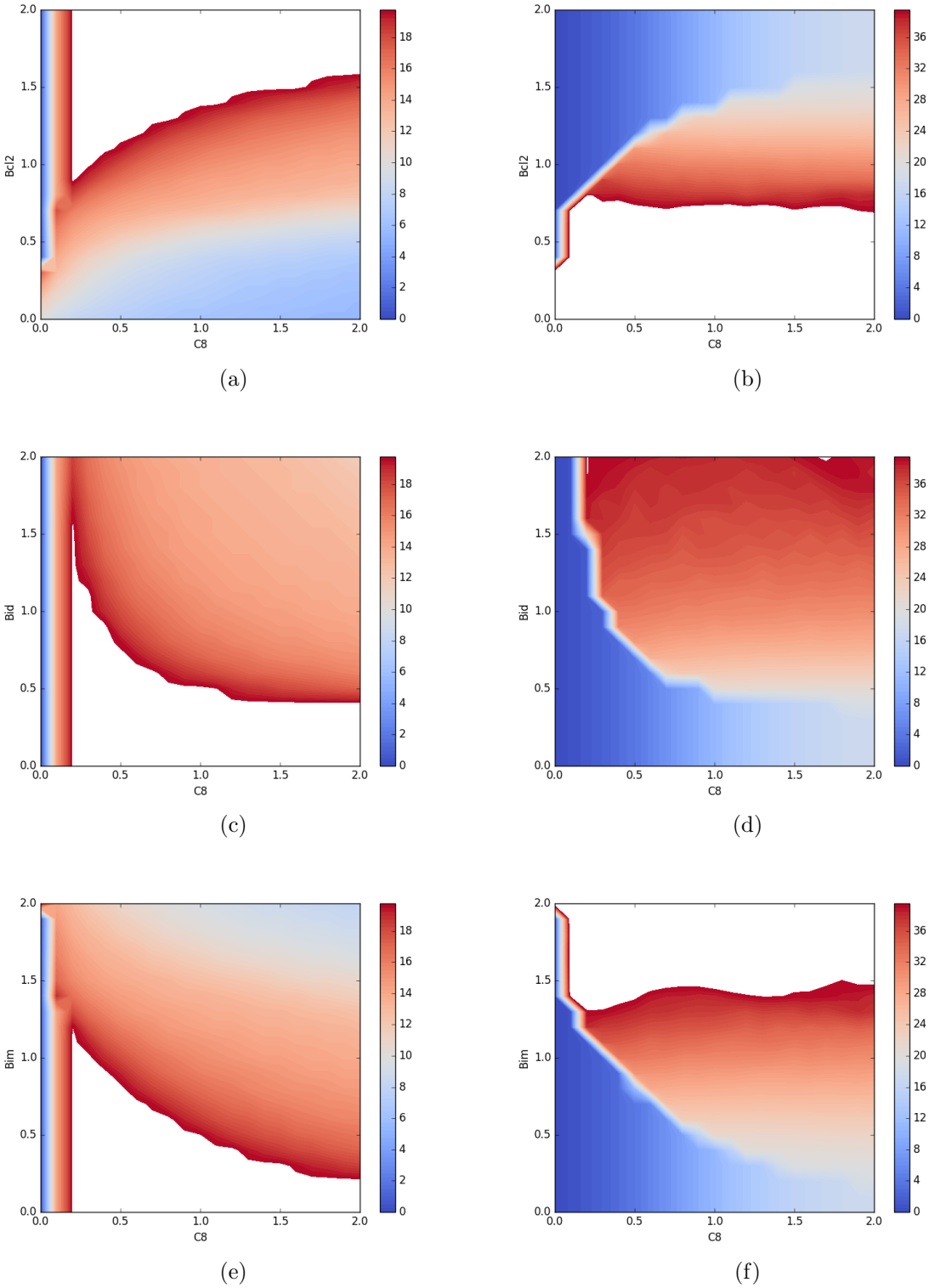
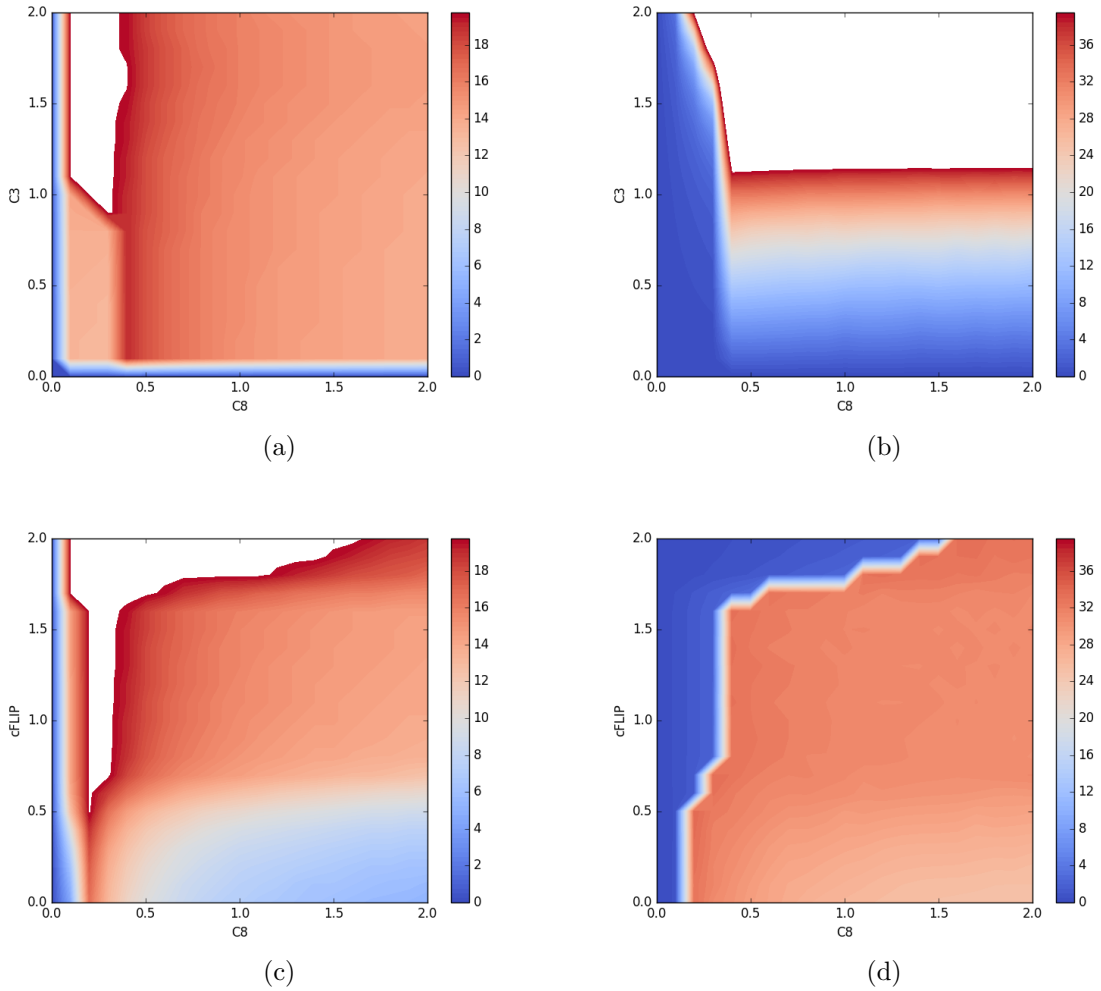


Figure E.7: Left: time to maximum concentration of activated Caspase 3 as two initial concentrations are perturbed. Right: Maximum concentration of activated Caspase 3 within 20 hours as the same initial concentrations are perturbed. The white indicates where the value exceeded the limit of the scale.



(a) (b)

(c) (d)

Figure E.8: Left: time to maximum concentration of activated Caspase 3 as two initial concentrations are perturbed. Right: Maximum concentration of activated Caspase 3 within 20 hours as the same initial concentrations are perturbed. The white indicates where the value exceeded the limit of the scale.

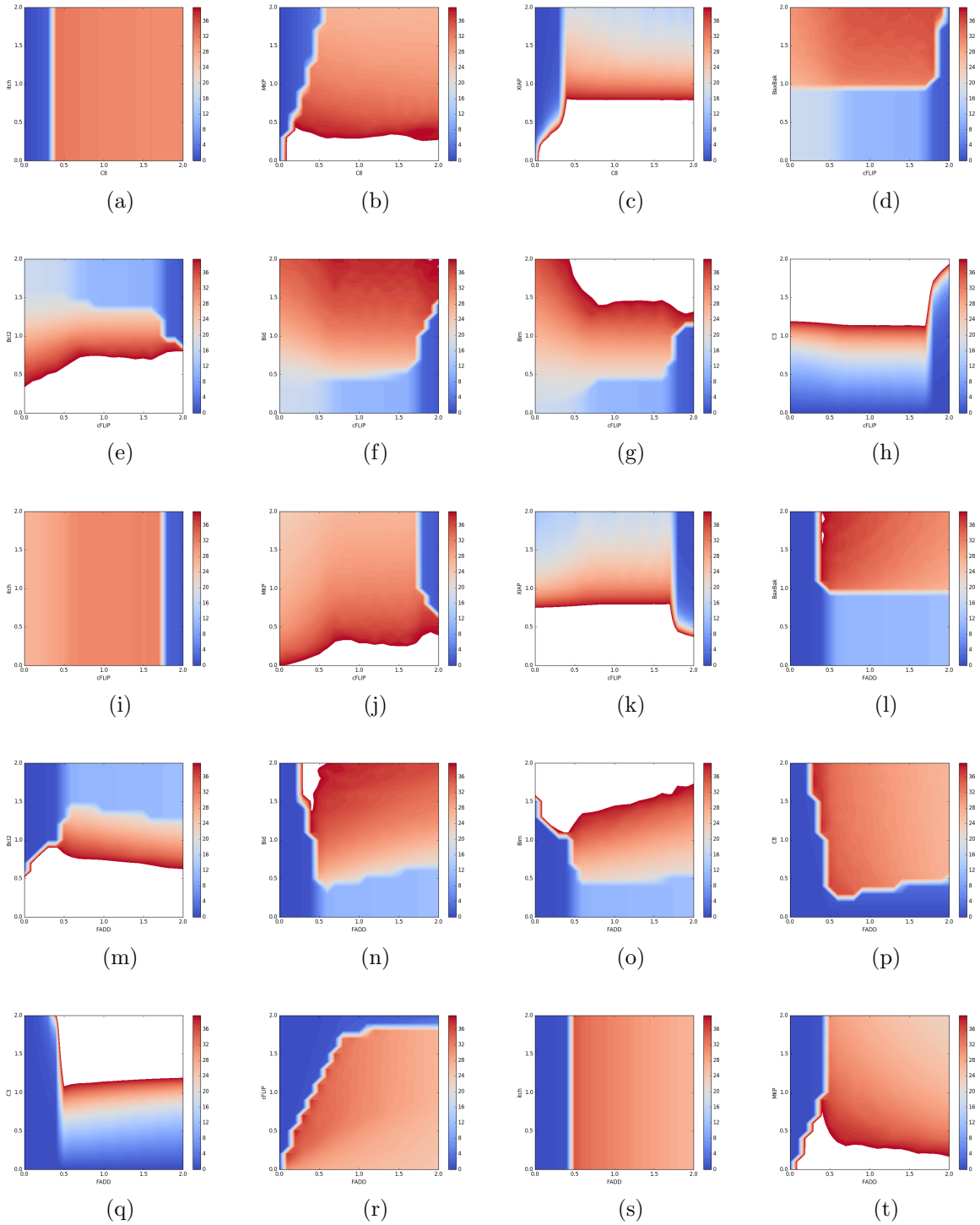


Figure E.9: Maximum concentration of activated Caspase 3 within 20 hours as the same initial concentrations are perturbed. The white indicates where the value exceeded the limit of the scale.

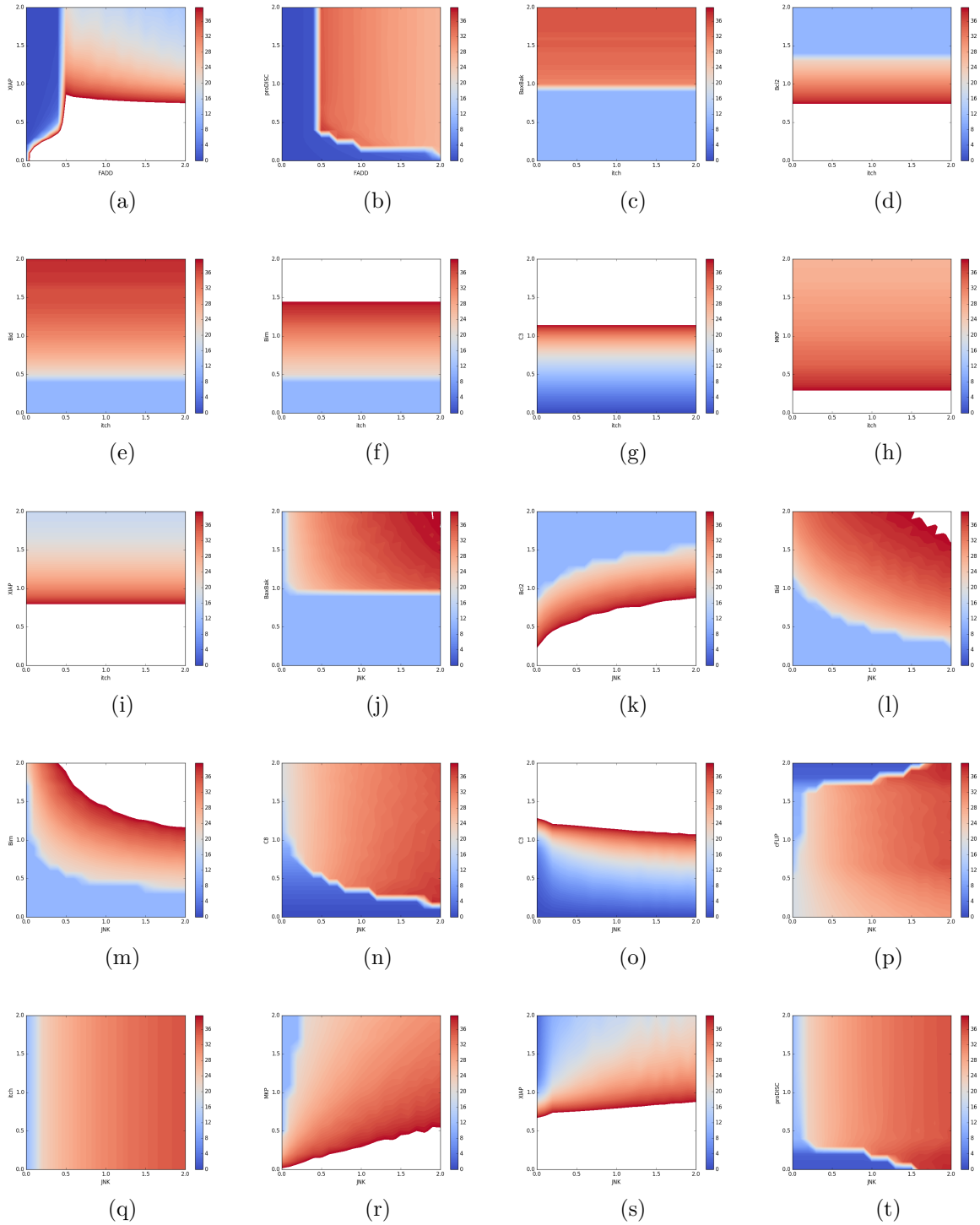


Figure E.10: Maximum concentration of activated Caspase 3 within 20 hours as the same initial concentrations are perturbed. The white indicates where the value exceeded the limit of the scale.

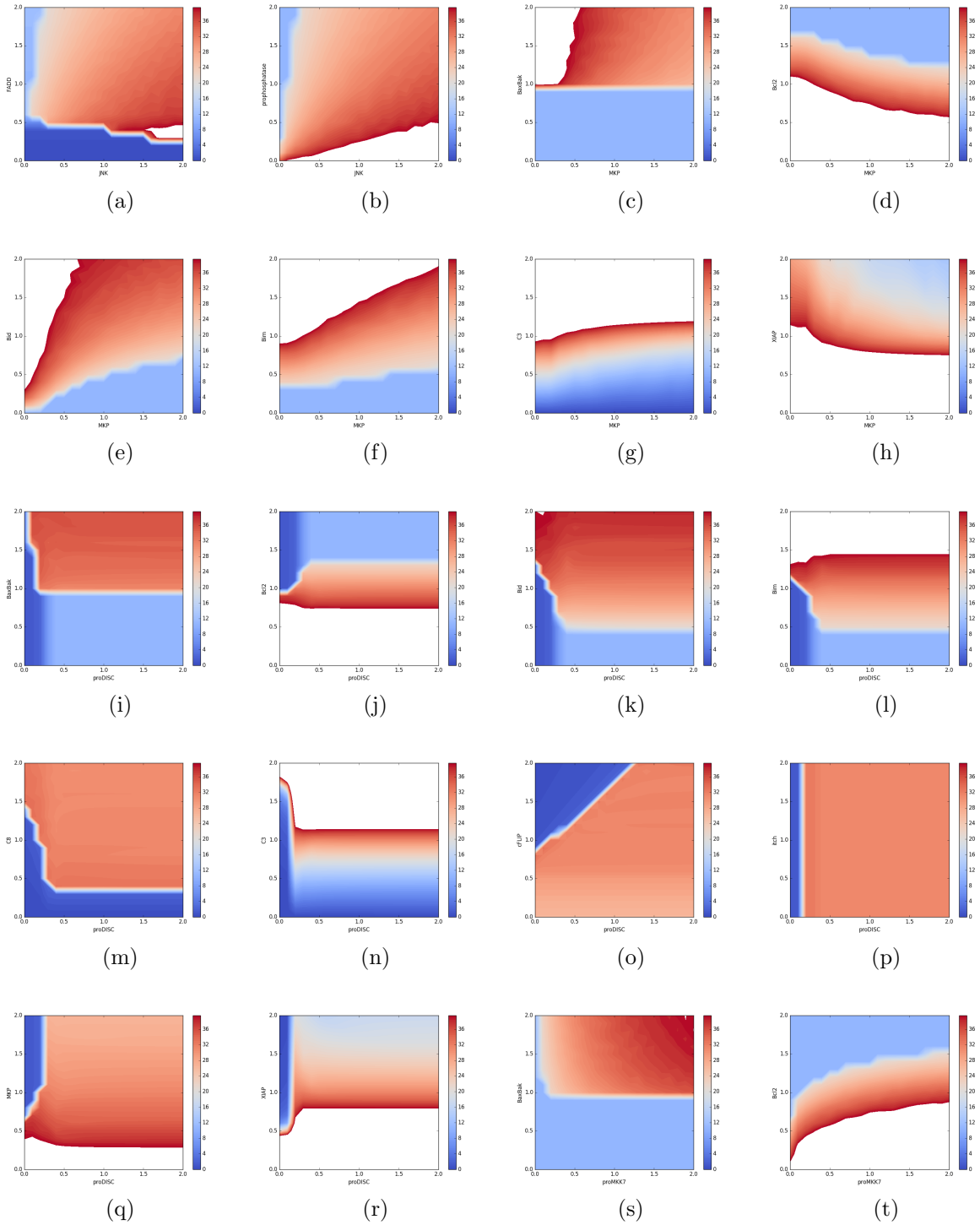


Figure E.11: Maximum concentration of activated Caspase 3 within 20 hours as the same initial concentrations are perturbed. The white indicates where the value exceeded the limit of the scale.

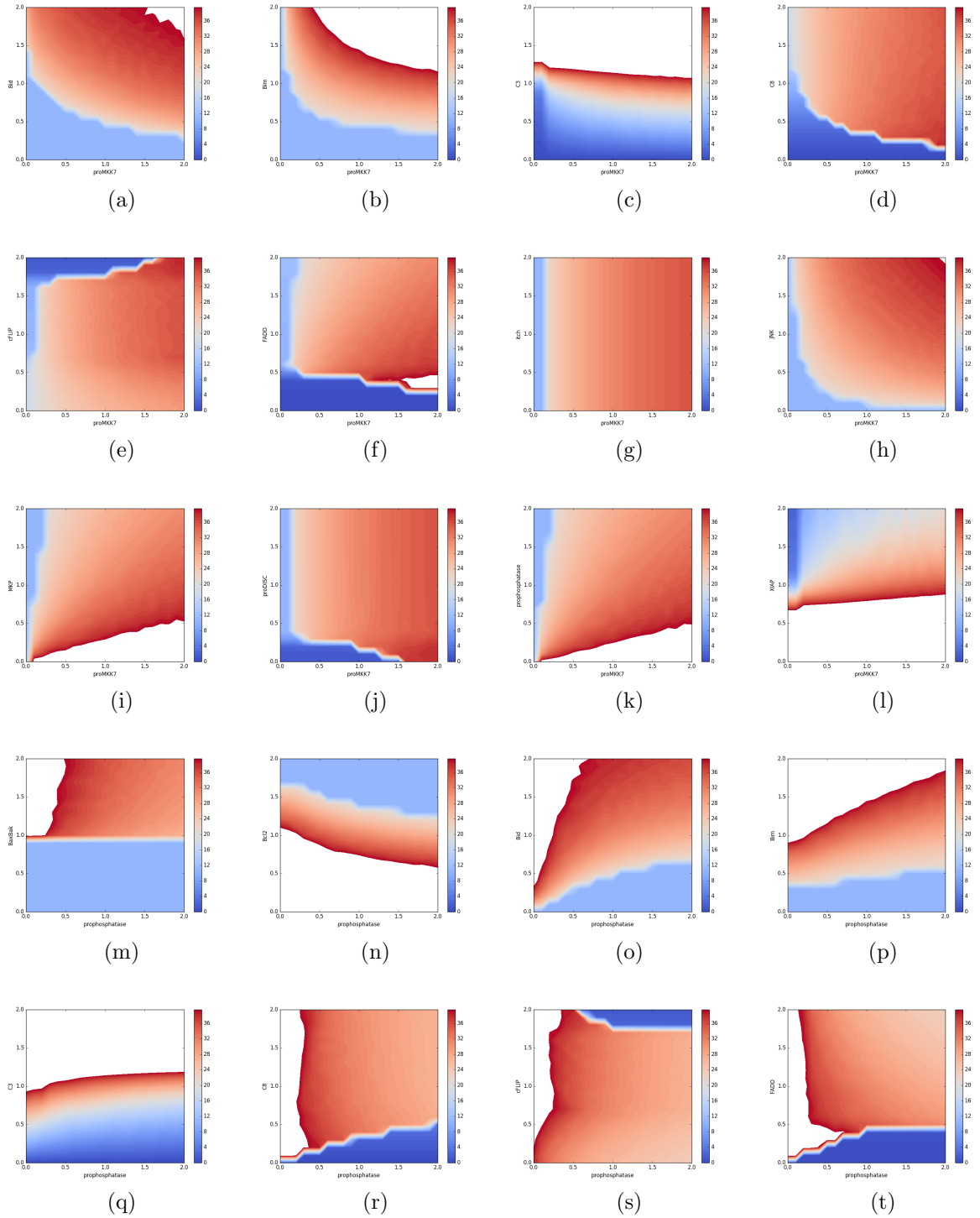


Figure E.12: Maximum concentration of activated Caspase 3 within 20 hours as the same initial concentrations are perturbed. The white indicates where the value exceeded the limit of the scale.

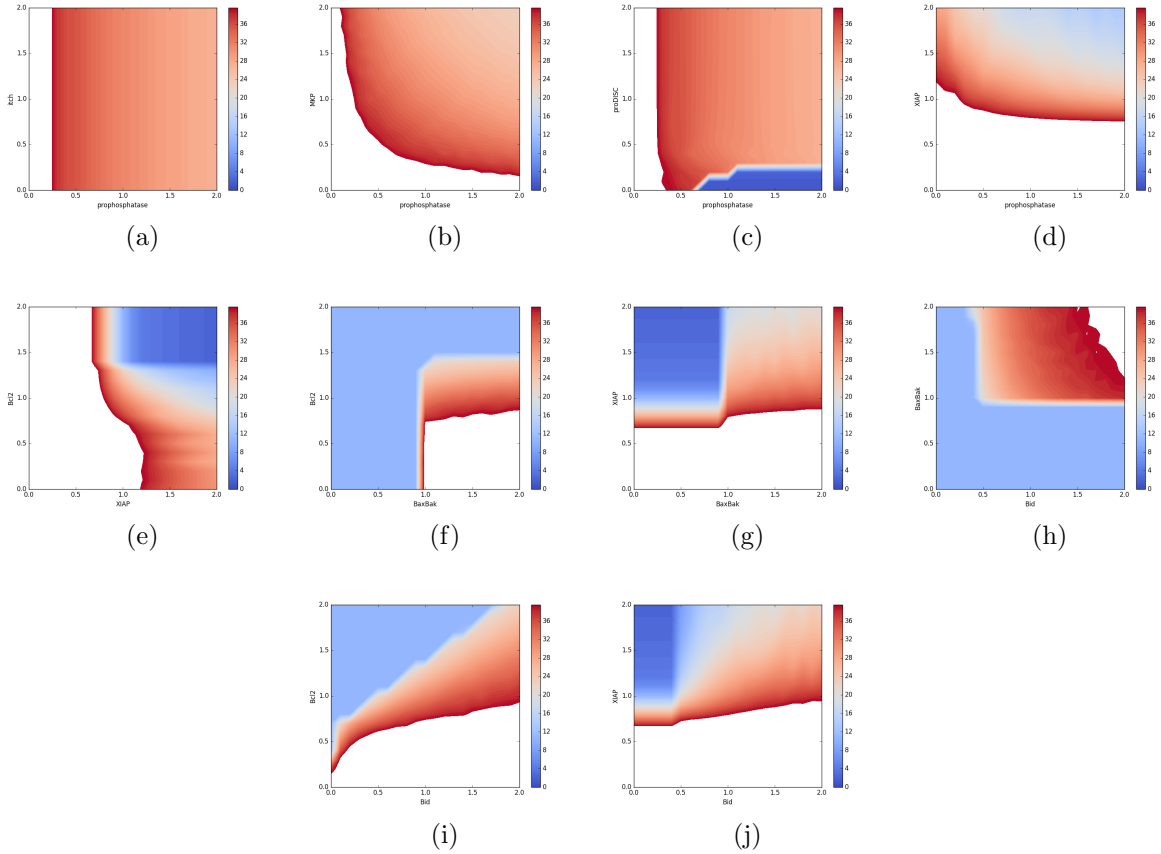


Figure E.13: Maximum concentration of activated Caspase 3 within 20 hours as the same initial concentrations are perturbed. The white indicates where the value exceeded the limit of the scale.

Appendix F

Simulated SNPs Used in the Separatrix Analysis

Table F.1: Simulated SNPs used in the smaller apoptosis model.

parameter	model effect	allele frequency	OR	p-value
k10neg	0.95	0.36	1.032	1.20359e-41
k10neg	0.95	0.26	1.03	2.49977e-31
k10neg	1.071	0.23	0.962	2.58248e-48
k10neg	0.93	0.21	1.047	2.18451e-63
k10neg	1.03	0.18	0.987	1.72789e-05
k10neg	0.954	0.12	1.027	1.84265e-14
k10neg	1.03	0.45	0.981	7.1697e-18
k10neg	0.917	0.45	1.055	1.01507e-127
k10neg	1.102	0.03	0.94	7.67061e-21
k10neg	0.964	0.4	1.019	1.88442e-16
k10neg	1.037	0.38	0.975	3.83447e-27
k10neg	0.983	0.27	1.011	5.6875e-05
k10neg	1.063	0.36	0.967	2.23132e-47
k10neg	1.08	0.03	0.951	3.46198e-14
k12neg	1.063	0.49	1.032	5.71594e-46
k12neg	1.048	0.15	1.027	3.78085e-17
k12neg	0.919	0.21	0.959	3.44214e-52
k12neg	0.95	0.47	0.977	1.29765e-25
k12neg	0.905	0.18	0.946	9.63293e-84
k12neg	0.908	0.09	0.947	4.40034e-44
k12neg	0.973	0.45	0.986	1.51826e-09
k12neg	1.073	0.16	1.039	1.33633e-35
k12neg	1.032	0.16	1.02	8.24874e-10
k12neg	1.109	0.19	1.057	5.37317e-86
k12neg	1.085	0.26	1.05	5.30302e-83

Table F.1: (continued)

k12neg	0.956	0.34	0.974	4.76778e-29
k12neg	1.022	0.18	1.014	4.06008e-06
k12neg	1.023	0.21	1.014	8.02893e-07
k12neg	1.032	0.07	1.019	7.84828e-05
k8neg	1.014	0.24	1.011	0.00014383
k8neg	0.957	0.43	0.972	6.54515e-36
k8neg	1.027	0.24	1.015	2.29851e-08
k8neg	1.037	0.25	1.023	9.93833e-18
k8neg	0.952	0.33	0.971	1.49403e-35
k8neg	0.965	0.44	0.975	1.73741e-28
k8neg	0.989	0.31	0.988	2.7434e-06
k8neg	0.96	0.13	0.972	1.92351e-17
k8neg	0.95	0.45	0.967	1.5715e-51
k8neg	1.066	0.34	1.04	2.52785e-63
k8neg	0.946	0.19	0.968	4.7838e-30
k9neg	1.051	0.25	0.976	1.24591e-20
k9neg	1.022	0.38	0.991	0.00049929
k9neg	1.153	0.2	0.931	6.51419e-149
k9neg	1.087	0.05	0.952	3.56676e-21
k9neg	0.977	0.11	1.017	9.60811e-06
k9neg	1.05	0.27	0.976	2.81066e-22
k9neg	0.915	0.46	1.05	5.56154e-108
k9neg	1.053	0.2	0.975	4.13315e-19
k9neg	0.961	0.2	1.024	1.07521e-16
k9neg	0.954	0.44	1.024	1.34972e-26

Table F.2: Simulated SNPs used in the larger apoptosis model.

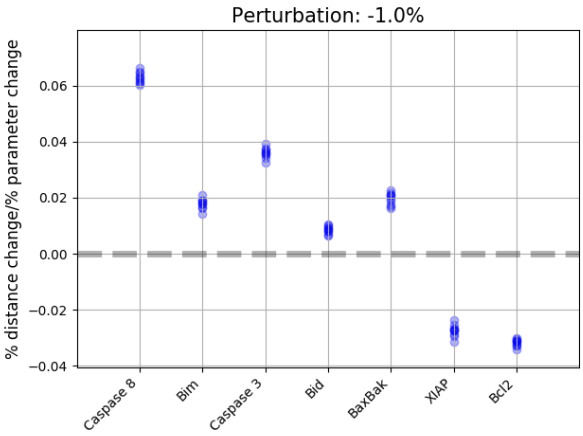
Gene	model effect	allele frequency	OR	p-value
BaxBak	1.122	0.42	0.967	1.11418e-08
BaxBak	1.168	0.42	0.958	2.23527e-13
Bcl2	1.095	0.13	1.113	9.63618e-38
Bcl2	1.14	0.08	1.158	5.38188e-45
Bcl2	1.095	0.31	1.123	1.2197e-81
Bcl2	1.062	0.21	1.059	4.48351e-16
Bcl2	0.936	0.04	0.923	1.34632e-07
Bcl2	1.063	0.26	1.072	6.73065e-27
Bcl2	1.148	0.36	1.13	4.19788e-96
Bcl2	1.078	0.18	1.07	2.63944e-19
Bcl2	1.077	0.15	1.095	4.57887e-31
Bcl2	0.902	0.33	0.902	3.10136e-67
Bid	0.793	0.49	1.109	1.43063e-75
Bid	1.191	0.12	0.904	6.85768e-31
Bid	0.909	0.41	1.024	0.000187298
Bid	0.921	0.38	1.034	5.3376e-08
Bid	1.053	0.26	0.97	1.25909e-05
Bim	0.962	0.35	1.025	0.000127028
Bim	0.881	0.18	1.078	1.12158e-24
Bim	0.867	0.24	1.076	1.3871e-28
Bim	0.934	0.02	1.062	0.00909134
Bim	0.905	0.16	1.061	3.26709e-14
Bim	0.84	0.36	1.095	1.08044e-53
Bim	0.943	0.39	1.027	1.82145e-05
Bim	0.935	0.1	1.061	9.24635e-10
C3	0.907	0.02	1.255	3.60061e-30
C3	1.088	0.29	0.829	4.19751e-202
C3	1.025	0.28	0.932	4.79703e-29
C3	0.928	0.3	1.153	6.36487e-119
C3	1.026	0.28	0.948	4.28202e-17
C8	0.88	0.23	1.024	0.00161818
C8	0.993	0.06	1.034	0.0181449
C8	0.815	0.09	1.038	0.000650635
FADD	1.135	0.41	1.017	0.0125757
JNK	1.22	0.41	0.96	4.17705e-12
JNK	0.851	0.45	1.032	9.84555e-08
JNK	0.827	0.08	1.032	0.00751134
XIAP	0.958	0.45	0.947	1.02727e-21
XIAP	0.946	0.01	0.923	0.0156286
XIAP	0.953	0.11	0.917	2.05165e-21
XIAP	0.964	0.45	0.951	1.26106e-18
XIAP	1.101	0.36	1.138	4.64115e-110

Table F.2: (continued)

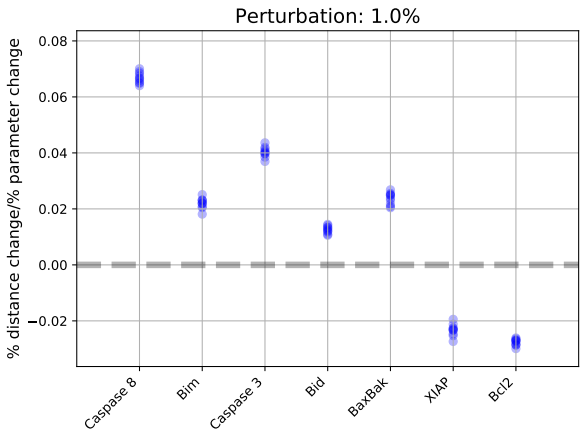
XIAP	1.151	0.27	1.225	2.57747e-233
XIAP	0.955	0.49	0.921	2.31732e-48
XIAP	0.911	0.05	0.862	2.88533e-29
XIAP	0.804	0.1	0.728	6.09411e-250
proMKK7	1.158	0.45	0.985	0.0223308
prophosphatase	0.866	0.44	0.98	0.00199392
prophosphatase	1.16	0.35	1.027	2.10535e-05
prophosphatase	0.814	0.06	0.955	0.000436115

Appendix G

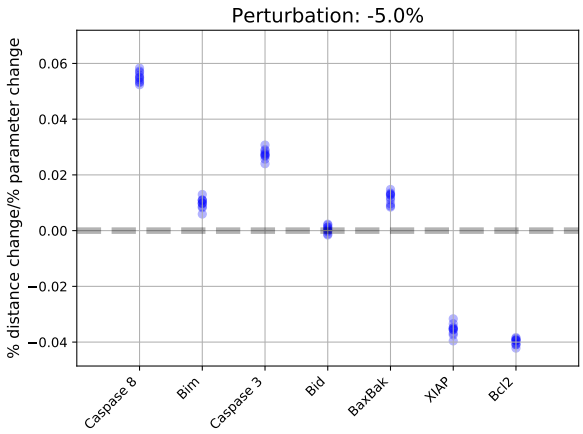
Separatrix Analysis of the Larger Apoptosis Model



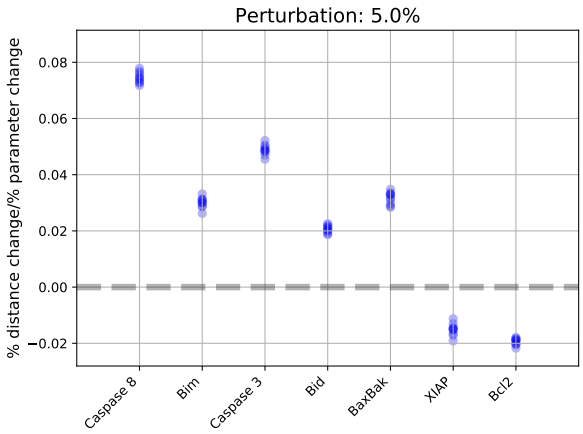
(a) -1% perturbation



(b) 1% perturbation

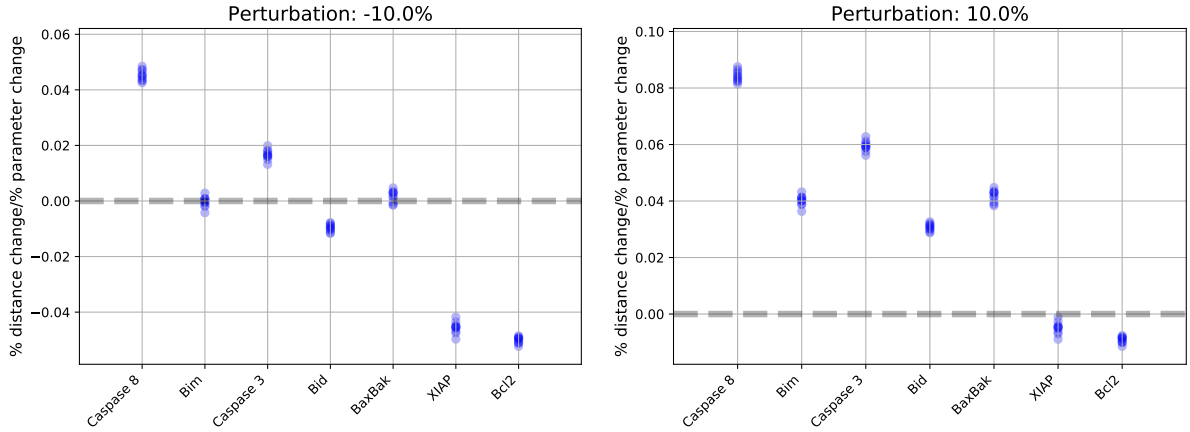


(c) -5% perturbation



(d) 5% perturbation

Figure G.1: Percentage mean distance change per percentage parameter change for each separatrix surface of the larger apoptosis model as each variable is perturbed one at a time, 0.01 (a-b) or 0.05 (c-d) times the initial value. (from left to right: proMKK7, JNK, prophosphatase, FADD, proDISC, Caspase 8, cFLIP, itch, MKP, Bim, Caspase 3, Bid, BaxBak, XIAP and Bcl2).



(a) -10% perturbation

(b) 10% perturbation

Figure G.2: Percentage mean distance change per percentage parameter change for each separatrix surface of the larger apoptosis model as each variable is perturbed one at a time, 0.01 (a-b) or 0.05 (c-d) times the initial value. (from left to right: proMKK7, JNK, prophosphatase, FADD, proDISC, Caspase 8, cFLIP, itch, MKP, Bim, Caspase 3, Bid, BaxBak, XIAP and Bcl2).

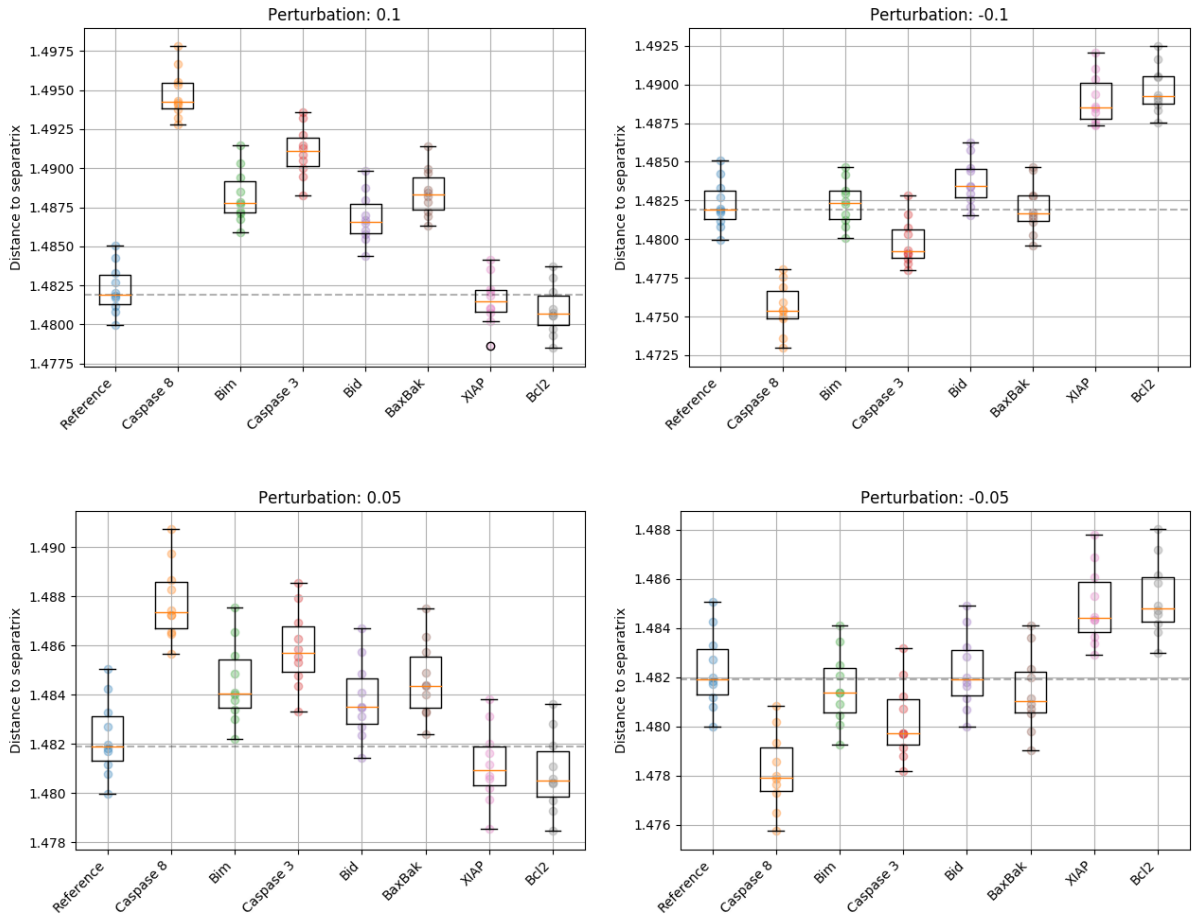
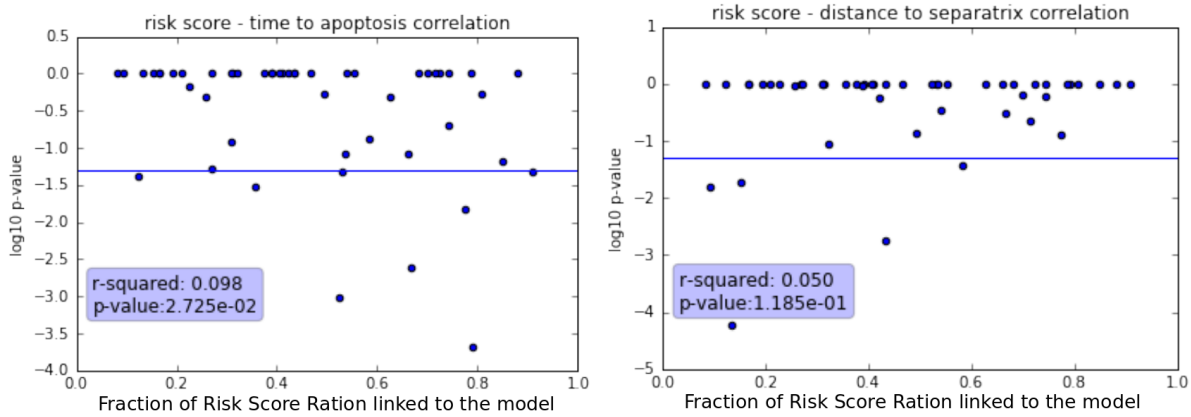
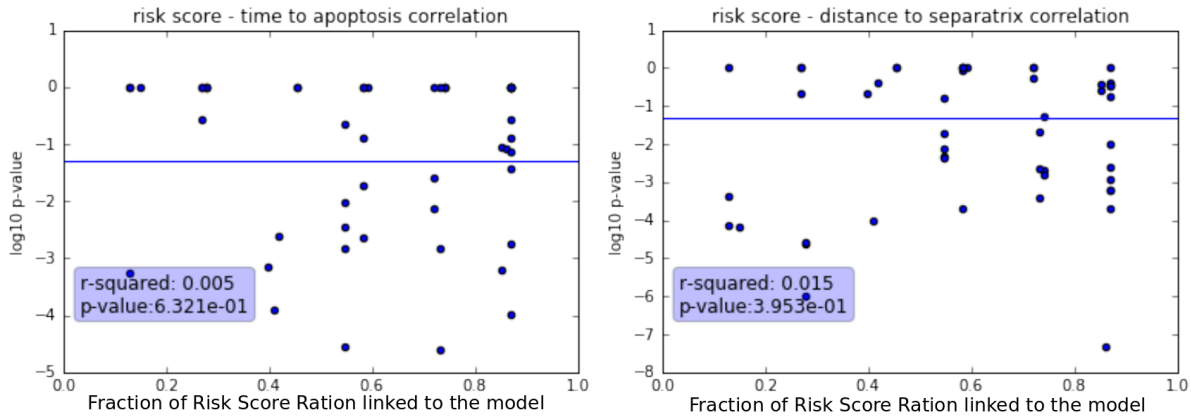


Figure G.3: Distance from starting point to separatrix surface for larger apoptosis model, calculated as mean distance to all points on the surface before and after each variable is perturbed 0.1, 0.05, -0.05 and -0.1 times the initial value. The first data indicate distances from original starting point to 10 surfaces and the rest indicate distances after perturbations or respective production rate parameter.



(a) time to apoptosis

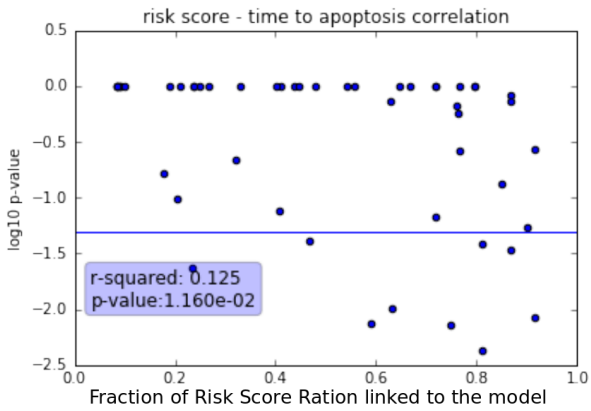
(b) distance to separatrix



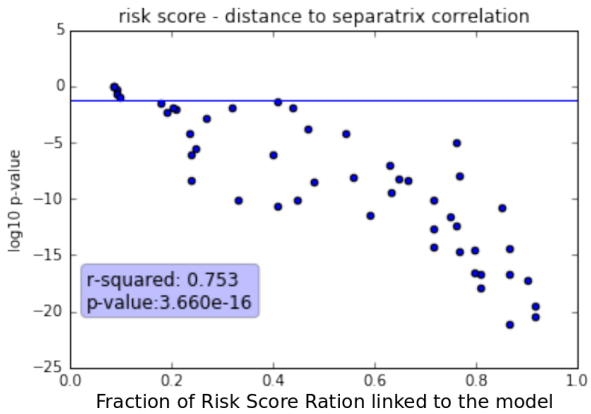
(c) time to apoptosis

(d) distance to separatrix

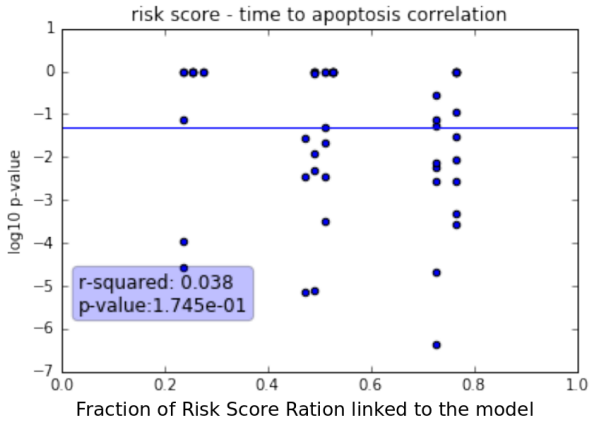
Figure G.4: 50 calculations of correlations between RSR and a: time to apoptosis and, b: distance to separatrix for the larger apoptosis model. Each simulations contained 50 individuals with a 10 SNPs randomly chosen from SNPs with a variable effect size between 0.96 and 1.06 and a random number of those SNPs linked to the model. Both maximum time to apoptosis and separatrix surface was set to 25 hours and only variables around the Caspase signalling were used.



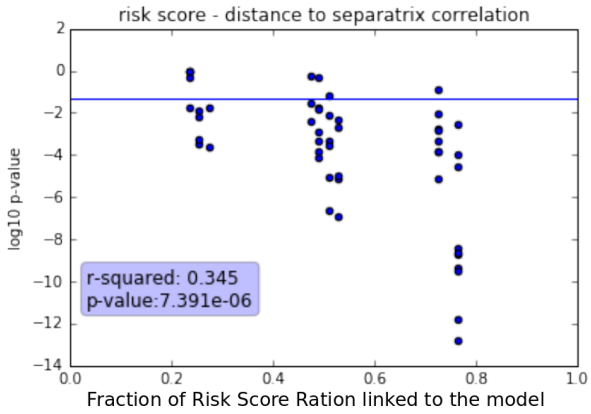
(a) time to apoptosis



(b) distance to separatrix

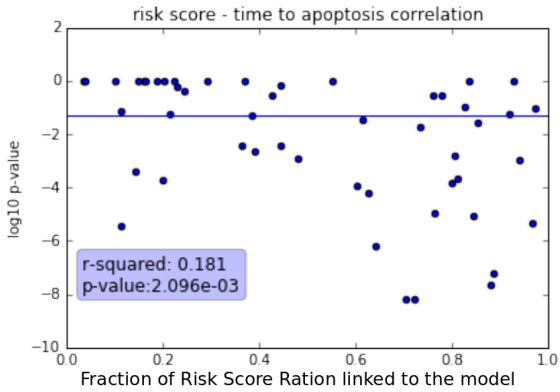


(c) time to apoptosis

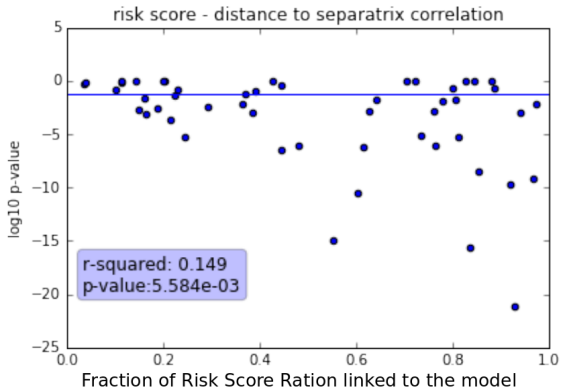


(d) distance to separatrix

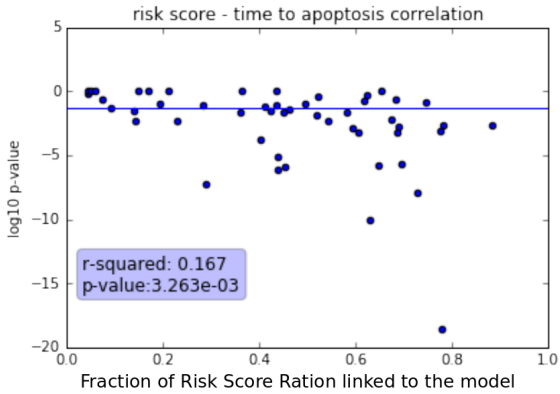
Figure G.5: 50 calculations of correlations between RSR and a: time to apoptosis and, b: distance to separatrix for the larger apoptosis model. Each simulations contained 50 individuals with a 10 SNPs randomly chosen from SNPs with a variable effect size between 0.96 and 1.06 and a random number of those SNPs linked to the model. Both maximum time to apoptosis and separatrix surface was set to 25 hours and only variables around the Caspase signalling were used.



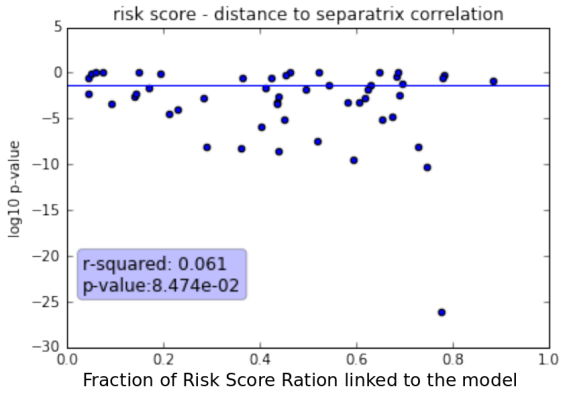
(a) time to apoptosis



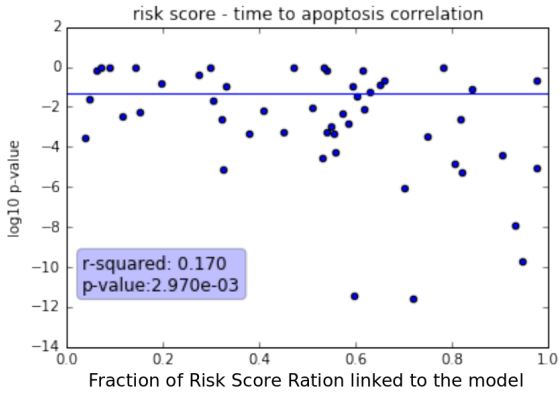
(b) distance to separatrix



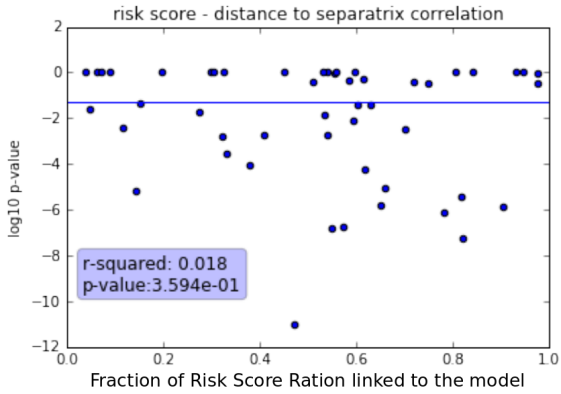
(c) time to apoptosis



(d) distance to separatrix

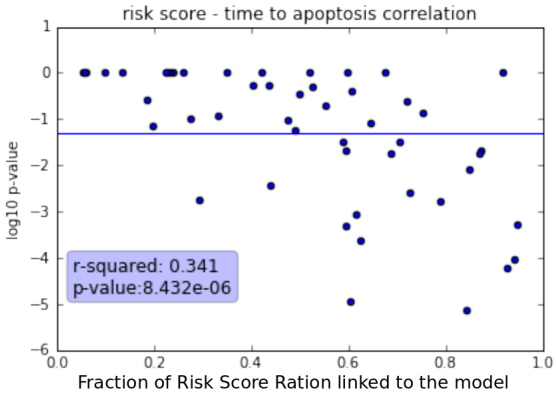


(e) time to apoptosis

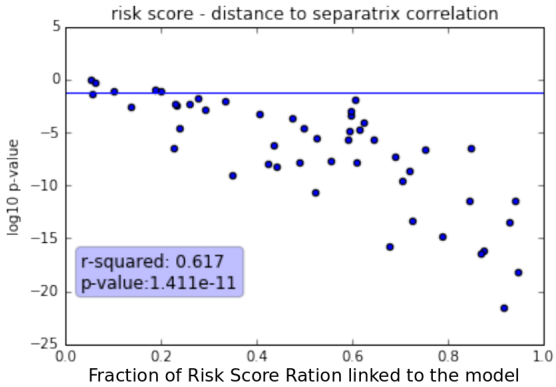


(f) distance to separatrix

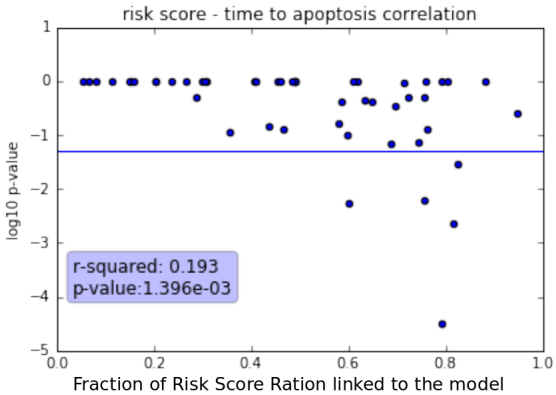
Figure G.6: 50 calculations of correlations between RSR and a,c and e: time to apoptosis and, b, d and f: distance to separatrix for the larger apoptosis model. Each simulations contained 50 individuals with a 10 SNPs randomly chosen from the entire data set of SNPs linked to included variables and a random number of those SNPs linked to the model. Both maximum time to apoptosis and separatrix surface was set to 25 hours (a–b), 50 hours (c–d) or 100 hours (e–f) and only variables around the Caspase signalling, excluding BaxBak, were used.



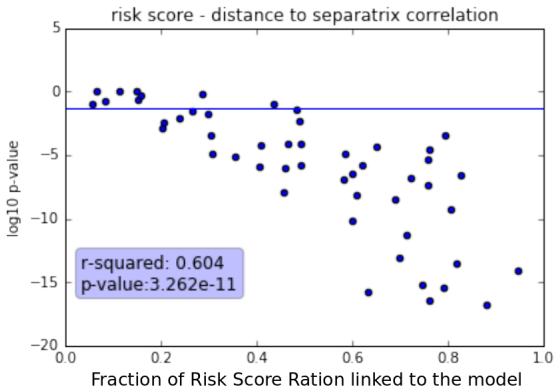
(a) time to apoptosis



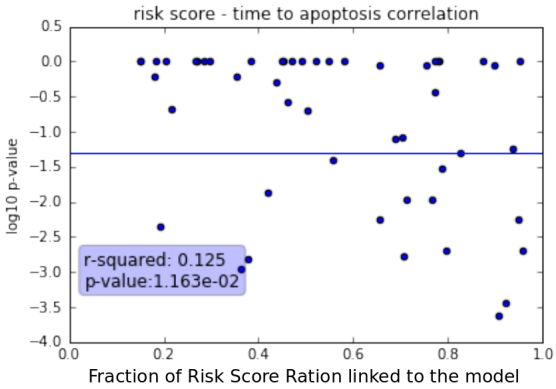
(b) distance to separatrix



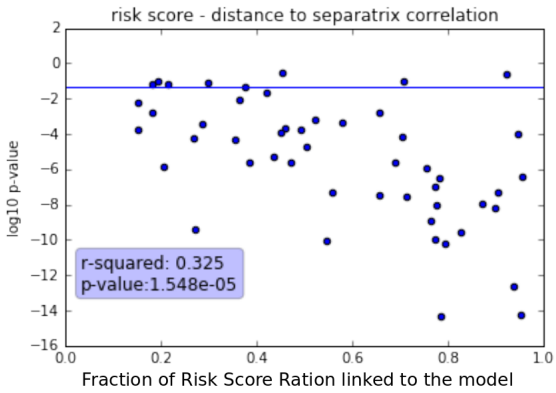
(c) time to apoptosis



(d) distance to separatrix



(e) time to apoptosis



(f) distance to separatrix

Figure G.7: 50 calculations of correlations between RSR and a, c and e: time to apoptosis and, b, d and f: distance to separatrix for the larger apoptosis model. Each simulations contained 50 individuals with a 10 SNPs randomly chosen from SNPs with a variable effect size between 0.97 and 1.03 and a random number of those SNPs linked to the model. Both maximum time to apoptosis and separatrix surface was set to 25 hours (a–b), 50 hours (c–d) or 100 hours (e–f) and only variables around the Caspase signalling, excluding BaxBak, were used.

Appendix H

Cancer associated SNPs

Table H.1: Breast cancer associated SNPs. When a closest gene was given in the literature, this is stated in the Gene column. If the SNP was used for the analysis using breast tissue and lymphoblastoid cell lines this is indicated by a 1 in the column breast and 1kG respectively

rs ID	oncoarray OR (95%CI)	Genes	breast	1kG
rs6596100	0.94(0.92–0.96)	HSPA4	1	1
rs79724016	0.93(0.88–0.97)	HIVEP3	1	1
rs6678914	1(0.99–1.02)	LGR6	1	1
rs11075995	1.03(1.01–1.06)	FTO	1	1
rs527616	0.97(0.95–0.98)	-	1	1
rs745570	1.03(1.01–1.05)	-	1	1
rs6562760	0.95(0.93–0.97)	-	1	1
rs7297051	0.89(0.87–0.91)	-	1	1
rs11820646	0.96(0.94–0.98)	-	1	1
rs3903072	0.97(0.95–0.99)	-	1	1
rs11199914	0.96(0.94–0.98)	-	1	1
rs13294895	1.06(1.03–1.08)	-	1	1
rs10816625	1.11(1.07–1.15)	-	1	1
rs10759243	1.06(1.04–1.08)	-	1	1
rs13281615	1.11(1.09–1.13)	-	1	1
rs13365225	0.91(0.89–0.93)	-	1	1
rs9693444	1.06(1.04–1.08)	-	1	1
rs17529111	1.02(1–1.04)	-	1	1
rs2012709	1.02(1–1.04)	-	1	1
rs4849887	0.91(0.88–0.94)	-	1	1
rs12710696	1.03(1.01–1.04)	-	1	1
rs4245739	1.02(1–1.04)	MDM4	1	1
rs1550623	0.95(0.93–0.98)	CDCA7	1	1
rs1432679	1.08(1.06–1.1)	EBF1	1	1
rs17356907	0.91(0.9–0.93)	NTN4	1	1

Table H.1: (continued)

rs4593472	0.97(0.95–0.99)	FLJ43663	1	1
rs6762644	1.05(1.03–1.07)	EGOT/ITPR1	1	1
rs12493607	1.05(1.03–1.07)	TGFBR2	1	1
rs6507583	0.92(0.89–0.96)	SETBP1	1	1
rs204247	1.04(1.02–1.06)	RANBP9	1	1
rs2943559	1.1(1.07–1.14)	HNF4G	1	1
rs10474352	0.94(0.92–0.97)	ARRDC3	1	1
rs4973768	1.11(1.09–1.13)	SLC4A7	1	1
rs3817198	1.05(1.03–1.07)	LSP1	1	1
rs941764	1.03(1.02–1.05)	CCDC88C	1	1
rs4577244	1.01(0.99–1.03)	WDR43	1	1
rs7904519	1.03(1.01–1.05)	TCFL2	1	1
rs2290203	0.94(0.92–0.96)	PRC1	1	1
rs1436904	0.95(0.94–0.97)	CHST9	1	1
rs11627032	0.96(0.94–0.98)	RIN3	1	1
rs7707921	0.96(0.94–0.98)	ATG10	1	1
rs6964587	1.03(1.02–1.05)	AKAP9	1	1
rs6828523	0.91(0.88–0.93)	ADAM29	1	1
rs1053338	1.05(1.02–1.07)	ATNX7	1	1
rs704010	1.07(1.05–1.09)	ZMZ1	1	1
rs4808801	0.93(0.91–0.95)	ELL	1	1
rs17426269	1.05(1.02–1.07)	-	1	1
rs16991615	1.1(1.06–1.14)	MCM8	1	1
rs2594714	0.97(0.95–0.99)	-	1	1
rs4496150	0.96(0.94–0.98)	-	1	1
rs206966	1.05(1.02–1.07)	-	1	1
rs17268829	1.05(1.03–1.07)	-	1	1
rs6815814	1.06(1.04–1.08)	-	1	1
rs12479355	0.96(0.94–0.98)	-	1	1
rs7529522	1.06(1.04–1.08)	-	1	1
rs12624860	1.04(1.01–1.07)	-	1	1
rs7971	0.96(0.94–0.98)	DNAH11, CDCA7L	1	1
rs310302	1.05(1.02–1.07)	-	1	1
rs1707302	0.96(0.95–0.98)	PIK3R3, LOC101929626	1	1
rs738321	0.95(0.93–0.97)	PLA2G6	1	1
rs11117758	0.95(0.93–0.97)	ESRRG	1	1
rs6805189	0.97(0.95–0.99)	FOXP1	1	1
rs2965183	1.04(1.02–1.06)	GATAD2A, MIR640	1	1
rs6569648	0.94(0.92–0.96)	L3MBTL3	1	1
rs10760444	1.03(1.02–1.05)	LMX1B	1	1
rs10022462	1.04(1.02–1.06)	LOC105369192	1	1

Table H.1: (continued)

rs9833888	1.06(1.04–1.08)	CMSS1, FILIP1L	1	1
rs16857609	1.06(1.04–1.09)	DIRC3	1	1
rs2981578	1.23(1.21–1.25)	FGFR2	1	1
rs11814448	1.12(1.06–1.19)	DNAJC1	1	1
rs7072776	1.05(1.03–1.07)	DNAJC1	1	1
rs999737	0.91(0.89–0.93)	RAD51B	1	1
rs2588809	1.06(1.03–1.08)	RAD51B	1	1
rs2380205	0.98(0.96–0.99)	ANKRD16	1	1
rs720475	0.96(0.94–0.98)	NOBOX, ARHGEF6	1	1
rs1292011	0.92(0.9–0.94)	TBX3	1	1
rs6001930	1.12(1.09–1.16)	MKL1	1	1
rs4784227	1.23(1.2–1.25)	TOX3	1	1
rs3819405	0.96(0.94–0.97)	ATXN1	1	1
rs13329835	1.07(1.05–1.09)	CDYL2	1	1
rs2823093	0.94(0.92–0.96)	NRIP1	1	1
rs11780156	1.05(1.03–1.08)	MYC	1	1
rs9397437	1.17(1.14–1.21)	ESR1	1	1
rs3757322	1.08(1.06–1.1)	ESR1	1	1
rs6725517	0.96(0.94–0.98)	ADCY3	1	1
rs116095464	1.06(1.02–1.1)	AHRR	1	1
rs1895062	0.94(0.92–0.95)	ASTN2	1	1
rs3760982	1.05(1.03–1.07)	KCCN4, LYPD5	1	1
rs17156577	1.05(1.02–1.08)	CREB5		1
rs34207738	1.06(1.04–1.08)	ZBTB38		1
rs78269692	1.09(1.04–1.13)	NFIX1		1
rs71801447	1.09(1.05–1.13)	BCL2L11		1
rs113577745	1.08(1.05–1.11)	GRHL1		1
rs151090251	1.10(1.05–1.16)	SMAD3		1
rs6597981	0.96(0.94–0.97)	PIDD1		1
rs35383942	1.12(1.08–1.17)	PHLDA3		1
rs6882649	0.97(0.95–0.99)	NREP		1
rs11389348	0.94(0.92–0.96)	-		1
rs6062356	1.09(1.06–1.12)	-		1
rs4233486	0.97(0.95–0.98)	-		1
rs140850326	0.97(0.95–0.99)	-		1
rs58058861	1.06(1.04–1.09)	-		1
rs77528541	0.95(0.92–0.97)	-		1
rs72749841	0.93(0.91–0.96)	-		1
rs35951924	0.95(0.93–0.97)	-		1
rs4562056	1.05(1.03–1.07)	-		1
rs71557345	0.92(0.88–0.96)	-		1

Table H.1: (continued)

rs12207986	0.97(0.95–0.98)	-		1
rs514192	1.05(1.03–1.07)	-		1
rs58847541	1.08(1.05–1.1)	-		1
rs67958007	1.09(1.06–1.12)	-		1
rs140936696	1.04(1.02–1.07)	-		1
rs202049448	0.95(0.93–0.97)	-		1
rs28539243	1.05(1.03–1.07)	-		1
rs6122906	1.05(1.03–1.07)	-		1
rs28512361	1.05(1.02–1.08)	-		1
rs2223621	1.04(1.02–1.06)	CDKAL1		1
rs9358466	0.96(0.94–0.98)	CASC15		1
rs71559437	0.93(0.91–0.96)	CUX1		1
rs2992756	1.06(1.04–1.08)	KLHDC7A		1
rs12546444	0.93(0.91–0.96)	ZFPM3		1
rs73161324	1.06(1.02–1.09)	XRCC6		1
rs4971059	1.05(1.03–1.07)	TRIM46		1
rs2432539	1.03(1.02–1.05)	AMFR		1
rs71338792	1.05(1.03–1.07)	GIPR		1
rs117618124	0.89(0.85–0.92)	GAREM1		1
rs13066793	0.94(0.91–0.97)	VGLL3		1
rs10623258	1.04(1.02–1.06)	ADSSL1		1
rs72826962	1.2(1.11–1.3)	CNTNAP1		1
rs1830298	1.06(1.04–1.08)	CASP8/ALS2CR12		1
rs2747652	0.94(0.92–0.96)	ESR1		1
rs35054928	1.27(1.25–1.3)	FGFR2		1
rs45631563	0.81(0.78–0.85)	FGFR2		1
rs11242675	1(0.98–1.02)	FOXQ1		1
rs2236007	0.93(0.91–0.95)	PAX9		1
rs12048493	1.04(1.02–1.06)	OTUD7B		1
rs11571833	1.35(1.23–1.48)	BRCA2		1
rs17879961	1.26(1.11–1.42)	CHEK2		1
rs75915166	1.28(1.24–1.33)	CCND1		1
rs34005590	0.82(0.79–0.86)	IGFBP5		1
rs132390	1.04(0.99–1.09)	EM1D1		1
rs1353747	0.96(0.93–0.99)	PDE4D		1
rs11552449	1.04(1.01–1.06)	DCLRE1B		1
rs6796502	0.92(0.89–0.95)	-		1
rs13162653	0.99(0.97–1.01)	-		1
rs10941679	1.15(1.13–1.18)	-		1
rs11977670	1.06(1.04–1.08)	-		1
rs6472903	0.94(0.92–0.96)	-		1

Table H.1: (continued)

rs12422552	1.06(1.04–1.08)	-		1
rs2787486	0.93(0.91–0.94)	-		1
rs67397200	1.03(1.01–1.05)	-		1
rs17817449	0.95(0.93–0.96)	FTO		1
rs12405132	0.97(0.95–0.99)	RNF115		1
rs9790517	1.04(1.01–1.06)	TET2		1
rs1011970	1.07(1.04–1.09)	CDKN2A, CDKN2B		1
rs62355902	1.18(1.15–1.21)	MAP3K1		1
rs72755295	1.15(1.09–1.2)	EXO1		1
rs10069690	1.06(1.04–1.08)	TERT		1
rs3215401	0.93(0.91–0.95)	TERT		1
rs10472076	1.03(1.01–1.04)	RAB3C		1
rs13267382	1.03(1.01–1.05)	LINC00536		1
rs2016394	0.95(0.94–0.97)	DLX2-AS1		1

Table H.2: Prostate cancer associated SNPs used in the analysis using prostate tissue. The SNP present in the array data and consequently used in the analysis in stated in the Proxy column. If any closest gene was given in the literature, this is stated in the Gene column.

SNP	Proxy	OR(OncoArray)	Gene
rs34925593	rs11691325	1.06	CDCA7
rs59308963	rs6754084	1.05	CASP8
rs1283104	rs1283102	1.04	DUBR
rs182314334	rs17370164	1.10	MBNL1
rs10793821	rs329114	1.05	RNU6-456P
rs9296068	rs9296068	1.05	HLA-DOA
rs9469899	rs2814971	1.05	UHRF1BP1
rs17621345	rs17621345	1.07	SUGCT
rs1048169	rs1048169	1.07	HAUS6
rs1182	rs2274507	1.07	TOR1A
rs141536087	rs12769002	1.10	LARP4B
rs7094871	rs7094463	1.04	TCF7L2
rs11290954	rs17749618	1.07	C11orf30, EMSY
rs1800057	rs1800056	1.13	ATM
rs878987	rs11223780	1.07	B3GAT1
rs10845938	rs10845938	1.06	RNU6-491P
rs7968403	rs2682714	1.07	RASSF3
rs5799921	rs10777195	1.08	RNU6-148P
rs11629412	rs910507	1.06	PAX9
rs4924487	rs4924490	1.06	CASC5
rs112293876	rs17851970	1.07	MAP2K1
rs12956892	rs8094161	1.05	OACYLP
rs10460109	rs7228257	1.04	TSHZ1
rs11666569	rs11666569	1.06	MYO9B
rs118005503	rs9304829	1.11	THEG5
rs11480453	rs4911110	1.05	DNMT3B
rs6091758	rs6126982	1.09	BCAS1
rs17599629	rs17599629	1.05	GOLPH3L
rs1218582	rs4845678	1.04	KCNN3
rs4245739	rs4245739	1.10	MDM4
rs9287719	rs6432112	1.08	-
rs1465618	rs4340576	1.09	THADA
rs721048	rs13417792	1.10	EHBP1
rs10187424	rs10198569	1.07	GGCX/VAMP8
rs12621278	rs16860397	1.26	ITGA6
rs3771570	rs17386695	1.09	FARP2
rs2660753	rs1865866	1.12	-
rs7611694	rs7611694	1.08	SIDT1
rs10934853	rs10934853	1.10	EEFSEC

Table H.2: (continued)

rs6763931	rs1344672	1.04	ZBTB38
rs1894292	rs1894292	1.05	AFM, RASSF6
rs12500426	rs7662466	1.05	PDLIM5
rs17021918	rs12500116	1.07	PDLIM5
rs7679673	rs10007915	1.11	TET2
rs12653946	rs10866528	1.08	IRX4
rs2121875	rs1482672	1.02	FGF10
rs3096702	rs9267873	1.05	NOTCH4
rs1983891	rs1983891	1.09	FOXP4
rs9443189	rs6906615	1.07	MYO6
rs2273669	rs6904998	1.06	ARMC2, SESN1
rs339331	rs339331	1.09	GPRC6A/RFX6
rs1933488	rs4083914	1.07	RSG17
rs9364554	rs12194182	1.10	SLC22A3
rs12155172	rs12155172	1.08	SP8
rs10486567	rs11982766	1.15	JAZF1
rs56232506	rs7801481	1.04	TNS3
rs6465657	rs11768309	1.10	LMTK2
rs2928679	rs2003976	1.05	SLC25A37
rs1512268	rs1160267	1.14	NKX3.1
rs11135910	rs6984769	1.06	EBF2
rs12543663	rs6984837	1.11	-
rs10086908	rs7842175	1.12	-
rs16901979	rs16901949	1.51	-
rs620861	rs1668875	1.16	-
rs6983267	rs6983267	1.21	-
rs1447295	rs7814837	1.43	-
rs17694493	rs17694493	1.05	CDKN2B-AS1
rs3850699	rs7904396	1.06	TRIM8
rs4962416	rs12769019	1.06	CTBP2
rs7931342	rs7109672	1.16	-
rs80130819	rs17122571	1.10	RP1-228P16.4
rs902774	rs902774	1.13	KRT8
rs1270884	rs1920568	1.07	TBX5
rs8008270	rs4901309	1.07	FERMT2
rs684232	rs1833459	1.09	VPS53, FAM57A
rs11649743	rs11649743	1.13	HNF1B
rs4430796	rs11651755	1.22	HNF1B
rs11650494	rs7216993	1.09	HOXB13, PRAX, SPOP, ZNF652
rs1859962	rs8072254	1.17	-
rs7241993	rs4799269	1.08	SALL3
rs8102476	rs7250689	1.11	-
rs11672691	rs2191139	1.09	-

Table H.2: (continued)

rs12480328	rs6091237	1.11	ADNP
rs6062509	rs6011040	1.06	ZGPAT
rs1041449	rs2838053	1.04	TMPRSS2
rs5759167	rs5759167	1.17	BIL/TTLL1
rs2405942	rs6530331	1.04	SHROOM2
rs5945619	rs1891702	1.11	NUDT11
rs2807031	rs4291439	1.06	-
rs5919432	rs4827556	1.04	AR
rs34925593	rs11691325	1.06	CDCA7
rs59308963	rs6754084	1.05	CASP8
rs1283104	rs1283102	1.04	DUBR
rs182314334	rs17370164	1.10	MBNL1
rs10793821	rs329114	1.05	RNU6-456P
rs9296068	rs9296068	1.05	HLA-DOA
rs9469899	rs2814971	1.05	UHRF1BP1
rs17621345	rs17621345	1.07	SUGCT
rs1048169	rs1048169	1.07	HAUS6
rs1182	rs2274507	1.07	TOR1A
rs141536087	rs12769002	1.10	LARP4B
rs7094871	rs7094463	1.04	TCF7L2
rs11290954	rs17749618	1.07	C11orf30, EMSY
rs1800057	rs1800056	1.13	ATM
rs878987	rs11223780	1.07	B3GAT1
rs10845938	rs10845938	1.06	RNU6-491P
rs7968403	rs2682714	1.07	RASSF3
rs5799921	rs10777195	1.08	RNU6-148P
rs11629412	rs910507	1.06	PAX9
rs4924487	rs4924490	1.06	CASC5
rs112293876	rs17851970	1.07	MAP2K1
rs12956892	rs8094161	1.05	OACYLP
rs10460109	rs7228257	1.04	TSHZ1
rs11666569	rs11666569	1.06	MYO9B
rs118005503	rs9304829	1.11	THEG5
rs11480453	rs4911110	1.05	DNMT3B
rs6091758	rs6126982	1.09	BCAS1
rs17599629	rs17599629	1.05	GOLPH3L
rs1218582	rs4845678	1.04	KCNN3
rs4245739	rs4245739	1.10	MDM4
rs9287719	rs6432112	1.08	-
rs1465618	rs4340576	1.09	THADA
rs721048	rs13417792	1.10	EHBP1
rs10187424	rs10198569	1.07	GGCX/VAMP8
rs12621278	rs16860397	1.26	ITGA6

Table H.2: (continued)

rs3771570	rs17386695	1.09	FARP2
rs2660753	rs1865866	1.12	-
rs7611694	rs7611694	1.08	SIDT1
rs10934853	rs10934853	1.10	EEFSEC
rs6763931	rs1344672	1.04	ZBTB38
rs1894292	rs1894292	1.05	AFM, RASSF6
rs12500426	rs7662466	1.05	PDLIM5
rs17021918	rs12500116	1.07	PDLIM5
rs7679673	rs10007915	1.11	TET2
rs12653946	rs10866528	1.08	IRX4
rs2121875	rs1482672	1.02	FGF10
rs3096702	rs9267873	1.05	NOTCH4
rs1983891	rs1983891	1.09	FOXP4
rs9443189	rs6906615	1.07	MYO6
rs2273669	rs6904998	1.06	ARMC2, SESN1
rs339331	rs339331	1.09	GPRC6A/RFX6
rs1933488	rs4083914	1.07	RSG17
rs9364554	rs12194182	1.10	SLC22A3
rs12155172	rs12155172	1.08	SP8
rs10486567	rs11982766	1.15	JAZF1
rs56232506	rs7801481	1.04	TNS3
rs6465657	rs11768309	1.10	LMTK2
rs2928679	rs2003976	1.05	SLC25A37
rs1512268	rs1160267	1.14	NKX3.1
rs11135910	rs6984769	1.06	EBF2
rs12543663	rs6984837	1.11	-
rs10086908	rs7842175	1.12	-
rs16901979	rs16901949	1.51	-
rs620861	rs1668875	1.16	-
rs6983267	rs6983267	1.21	-
rs1447295	rs7814837	1.43	-
rs17694493	rs17694493	1.05	CDKN2B-AS1
rs3850699	rs7904396	1.06	TRIM8
rs4962416	rs12769019	1.06	CTBP2
rs7931342	rs7109672	1.16	-
rs80130819	rs17122571	1.10	RP1-228P16.4
rs902774	rs902774	1.13	KRT8
rs1270884	rs1920568	1.07	TBX5
rs8008270	rs4901309	1.07	FERMT2
rs684232	rs1833459	1.09	VPS53, FAM57A
rs11649743	rs11649743	1.13	HNF1B
rs4430796	rs11651755	1.22	HNF1B
rs11650494	rs7216993	1.09	HOXB13, PRAX, SPOP, ZNF652

Table H.2: (continued)

rs1859962	rs8072254	1.17	-
rs7241993	rs4799269	1.08	SALL3
rs8102476	rs7250689	1.11	-
rs11672691	rs2191139	1.09	-
rs12480328	rs6091237	1.11	ADNP
rs6062509	rs6011040	1.06	ZGPAT
rs1041449	rs2838053	1.04	TMPRSS2
rs5759167	rs5759167	1.17	BIL/TTLL1
rs2405942	rs6530331	1.04	SHROOM2
rs5945619	rs1891702	1.11	NUDT11
rs2807031	rs4291439	1.06	-
rs5919432	rs4827556	1.04	AR

Appendix I

Separatrix Analysis Using Experimental Data

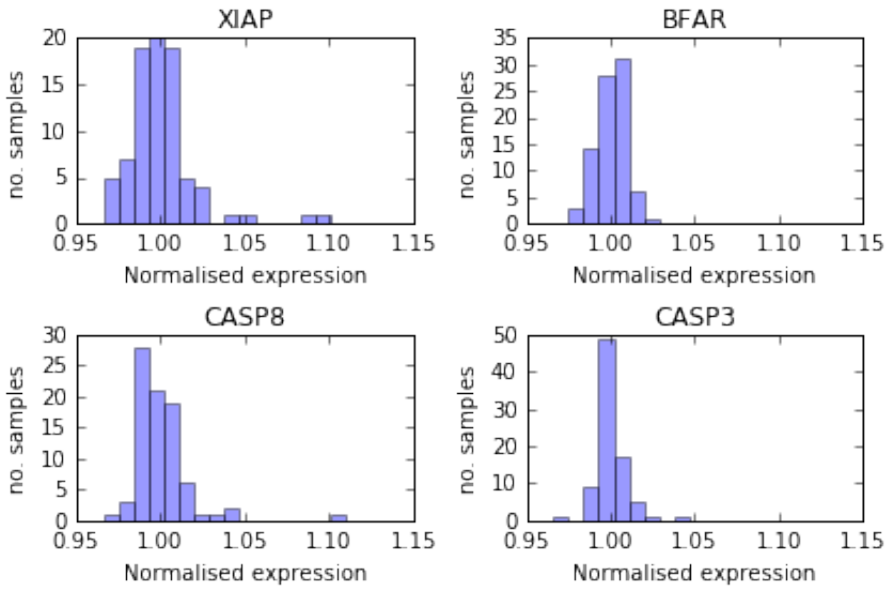


Figure I.1: Distribution of normalised RNA expression for the four genes XIAP, BFAR, CASP8 and CASP3 in lymphoblastoid cell lines from 1000 genome project. For each sample 95% of the difference between the expression and the mean expression was subtracted and all samples, were then divided by the new mean expression.

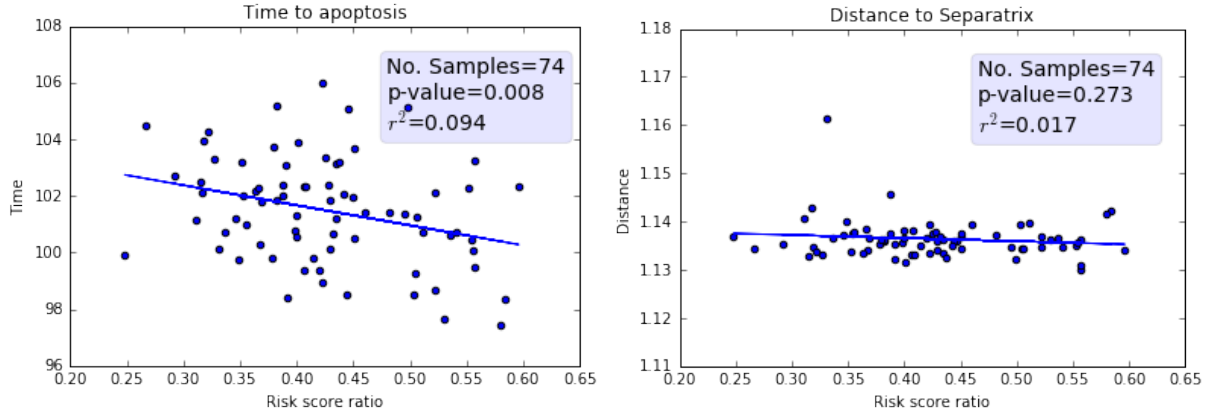


Figure I.2: Left: Correlation between RSR and time to apoptosis, as calculated by the smaller apoptosis model, for the lymphoblastoid cell lines from the 1000 genome project. Right: Correlation between RSR and distance to separatrix. In both cases the 12 previously selected SNPs were used to calculate the RSR and the 2 samples with an abnormal response time were excluded from the analysis.

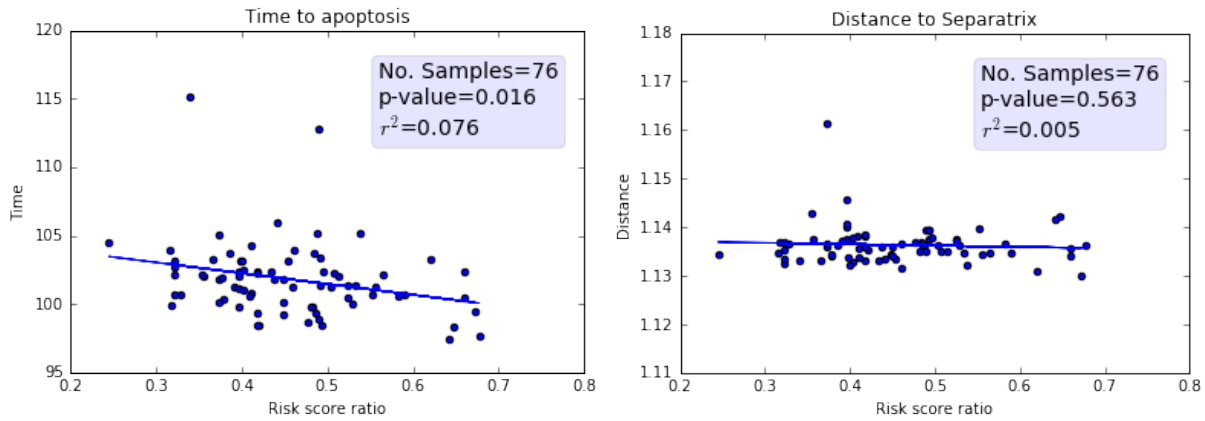


Figure I.3: Left: Correlation between RSR and time to apoptosis for the lymphoblastoid cell lines from the 1000 genome project. Right: Correlation between RSR and distance to separatrix. In both cases 9 SNPs thought to be linked to one of the genes in the model, which could also be used for the breast tissue later on, were used to calculate the RSR.

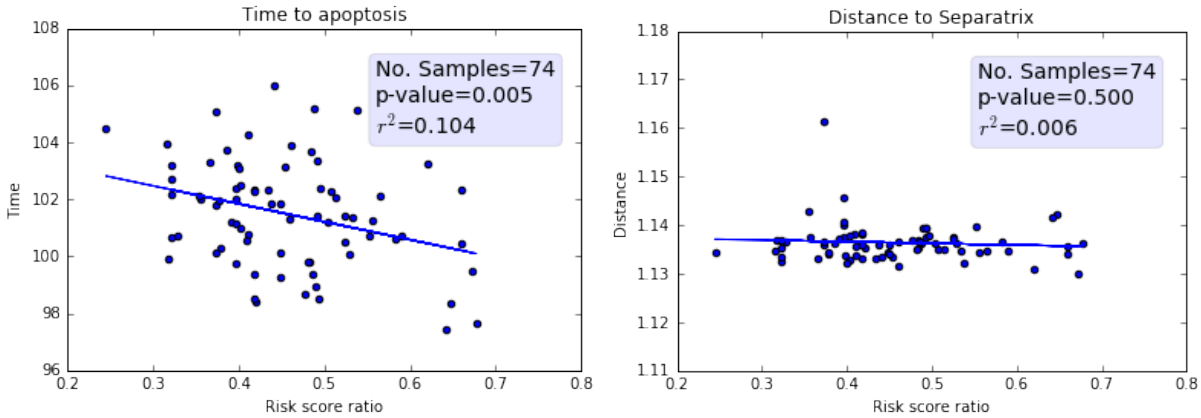


Figure I.4: Left: Correlation between RSR and time to apoptosis, as calculated by the smaller apoptosis model, for the lymphoblastoid cell lines from the 1000 genome project. Right: Correlation between RSR and distance to separatrix. In both cases the 9 previously selected SNPs, which could also be used for the breast tissue later on, were used to calculate the RSR and the 2 samples with an abnormal response time were excluded from the analysis.

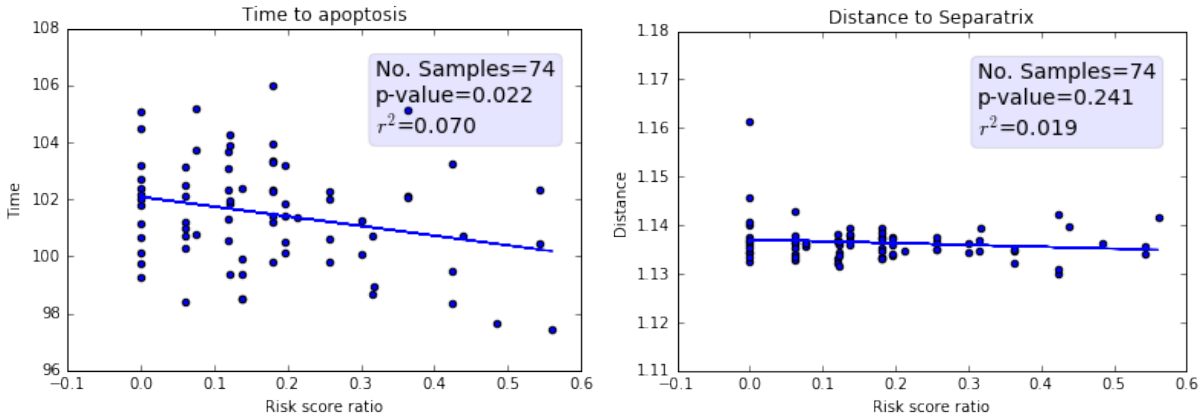


Figure I.5: Left: Correlation between RSR and time to apoptosis for the lymphoblastoid cell lines from the 1000 genome project. Right: correlation between risk score ratio and distance to apoptosis for the smaller apoptosis model. In both cases the 4 SNPs previously identified to map to proteins with binding sites in a promoter of one of the genes in the model were used to calculate the RSR and the 2 samples with an abnormal response time were excluded from the analysis.

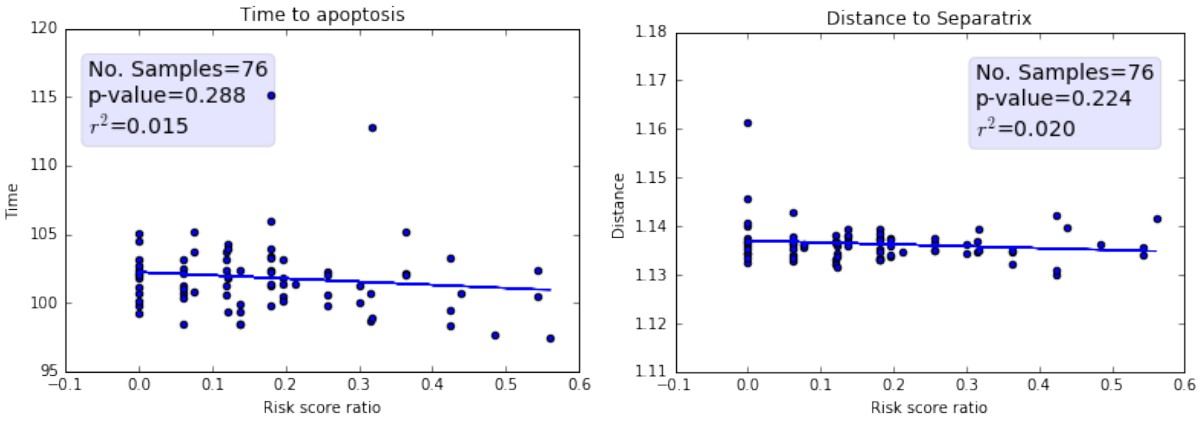


Figure I.6: Left: Correlation between RSR and time to apoptosis for the lymphoblastoid cell lines from the 1000 genome project. Right: correlation between RSR and distance to apoptosis for the smaller apoptosis model. Out of the 9 previously selected SNPs the 4 SNPs which mapped to transcription regulators with binding sites in the promoter of one of the four genes in the model were extracted and used to calculate the RSR.

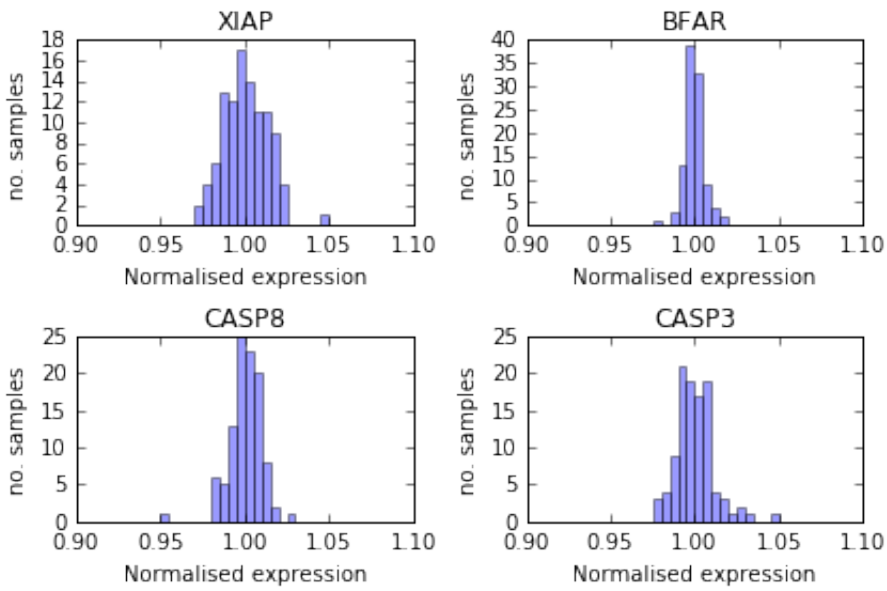


Figure I.7: Distribution of normalised RNA expression for the four genes XIAP, BFAR, CASP8 and CASP3 in normal breast tissue. For each sample 95% of the difference between the expression and the mean expression was subtracted and all samples, were then divided by the new mean expression.

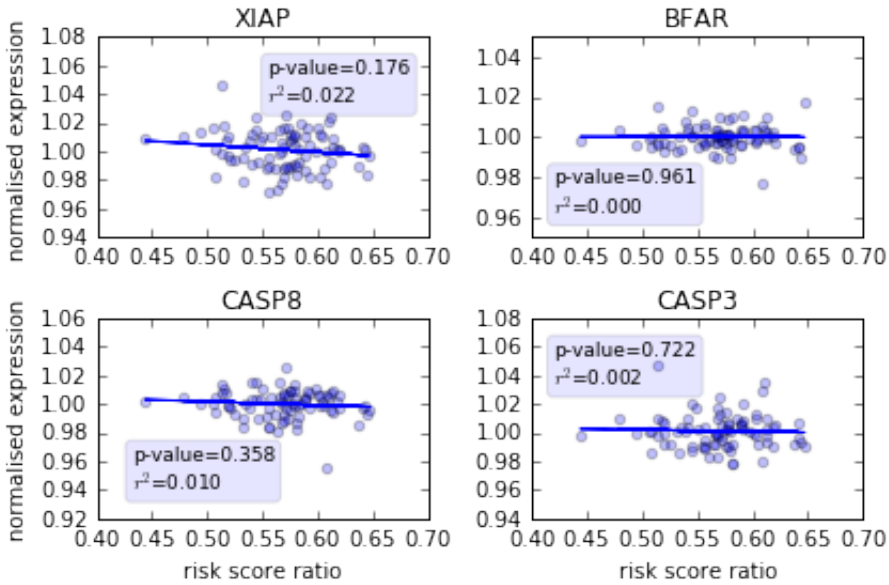


Figure I.8: Correlation between RSR on the x-axis and normalised RNA expression on the y-axis for the four genes XIAP, BFAR, CASP8 and CASP3 in normal breast tissue.

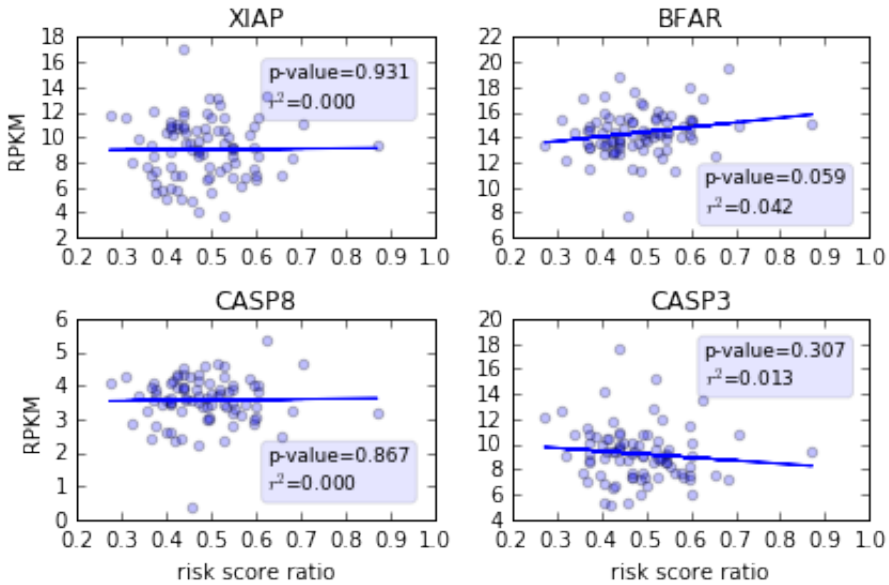


Figure I.9: Correlation between RSR of the 9 selected breast cancer associated SNPs on the x-axis and normalised RNA expression on the y-axis for the four genes XIAP, BFAR, CASP8 and CASP3 in normal breast tissue.

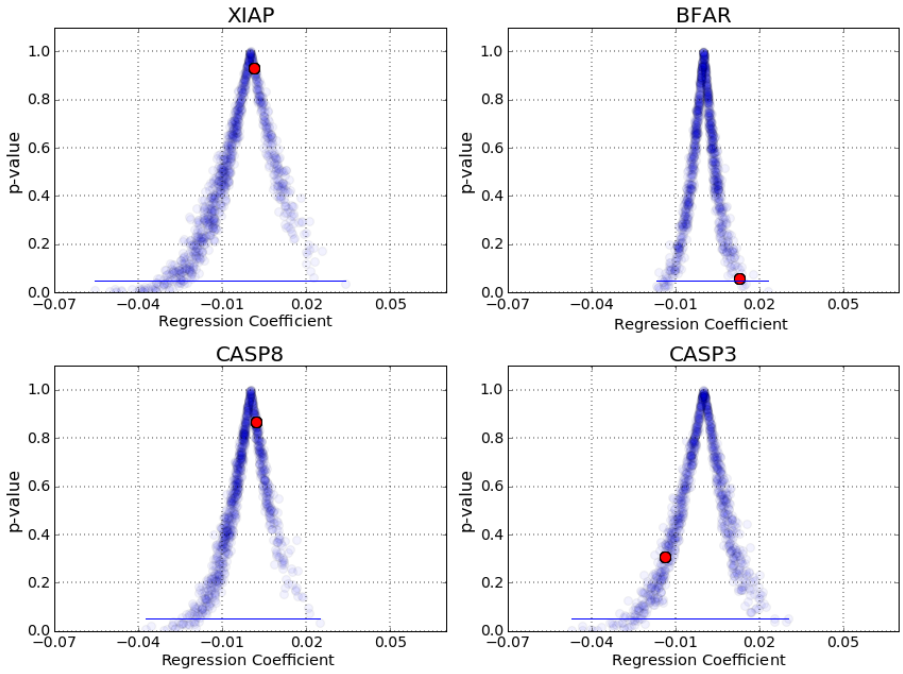


Figure I.10: Distribution of p-values and regression coefficients for correlations between RSR of 9 randomly selected SNPs and the expression values of the four genes XIAP, BFAR, CASP8 and CASP3 in normal breast tissue. The 9 SNPs previously selected for further study are marked in red.

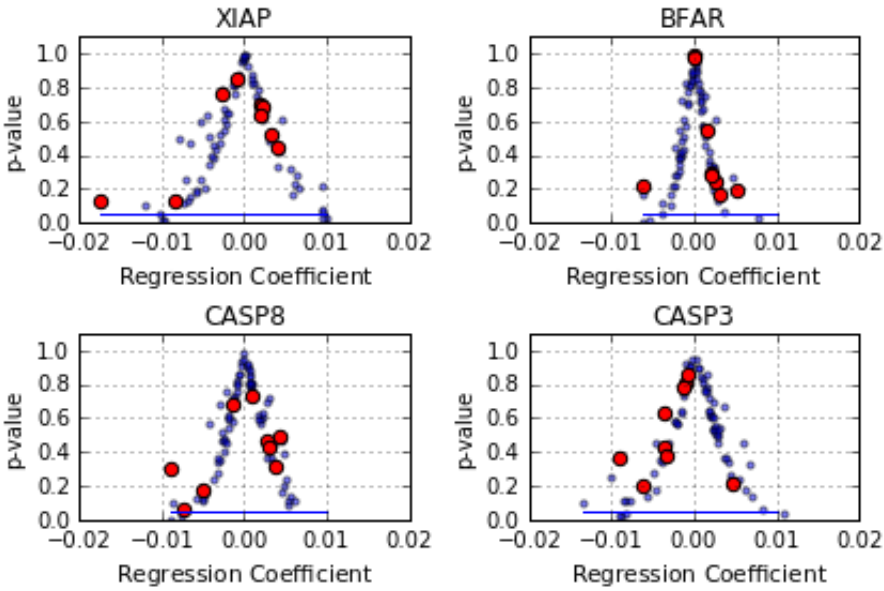


Figure I.11: Distribution of p-values and regression coefficients for correlations between RSR of one SNP at a time and the expression values for the four genes XIAP, BFAR, CASP8 and CASP3 in normal breast tissue. The 9 SNPs previously selected for further study are marked in red.

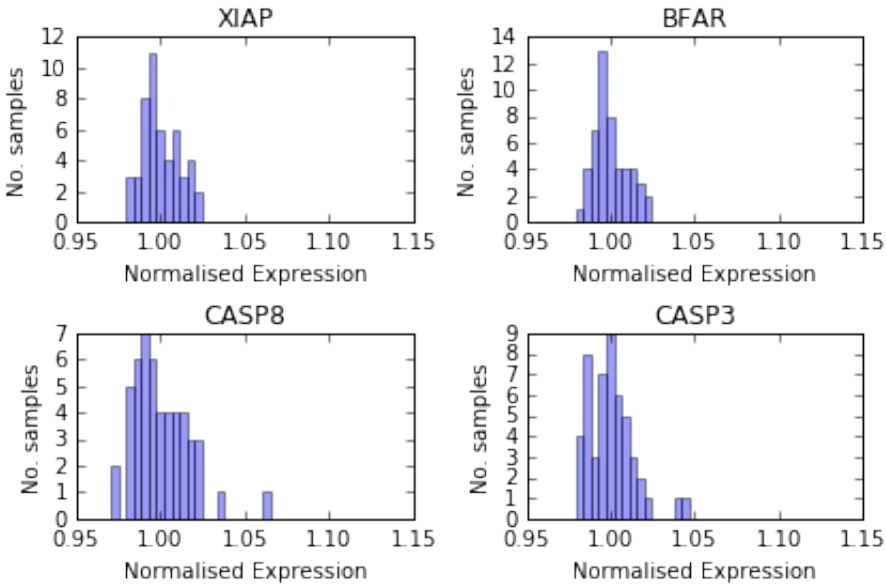


Figure I.12: Distribution of normalised RNA expression for the four genes XIAP, BFAR, CASP8 and CASP3 in normal prostate tissue. For each sample 95% of the difference between the expression and the mean expression was subtracted and all samples were then divided by the new mean expression.

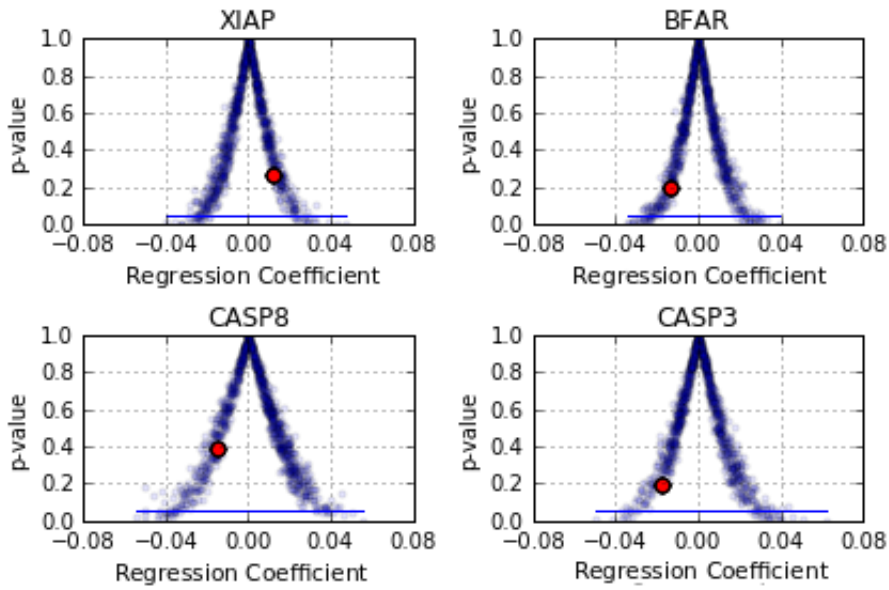


Figure I.13: Distribution of p-values and regression coefficients for correlations between RSR of 7 randomly selected SNPs and the expression values of the four genes XIAP, BFAR, CASP8 and CASP3 in normal prostate tissue. The 7 SNPs previously selected for further study are marked in red.

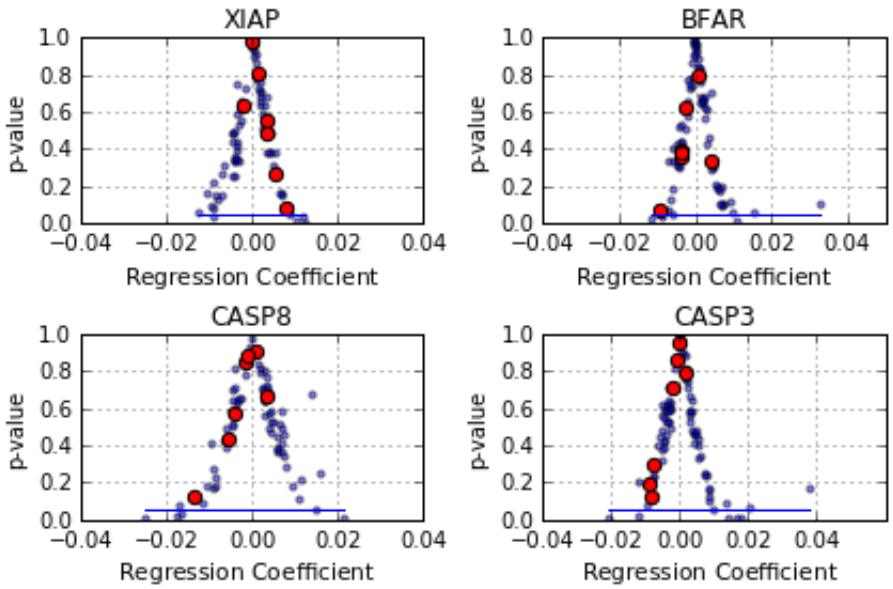


Figure I.14: Distribution of p-values and regression coefficients for correlations between RSR of single SNPs and the expression values for XIAP, BFAR, CASP8 and CASP3 in normal prostate tissue. The distribution of the correlations for 7 SNPs previously selected for further study are marked in red.

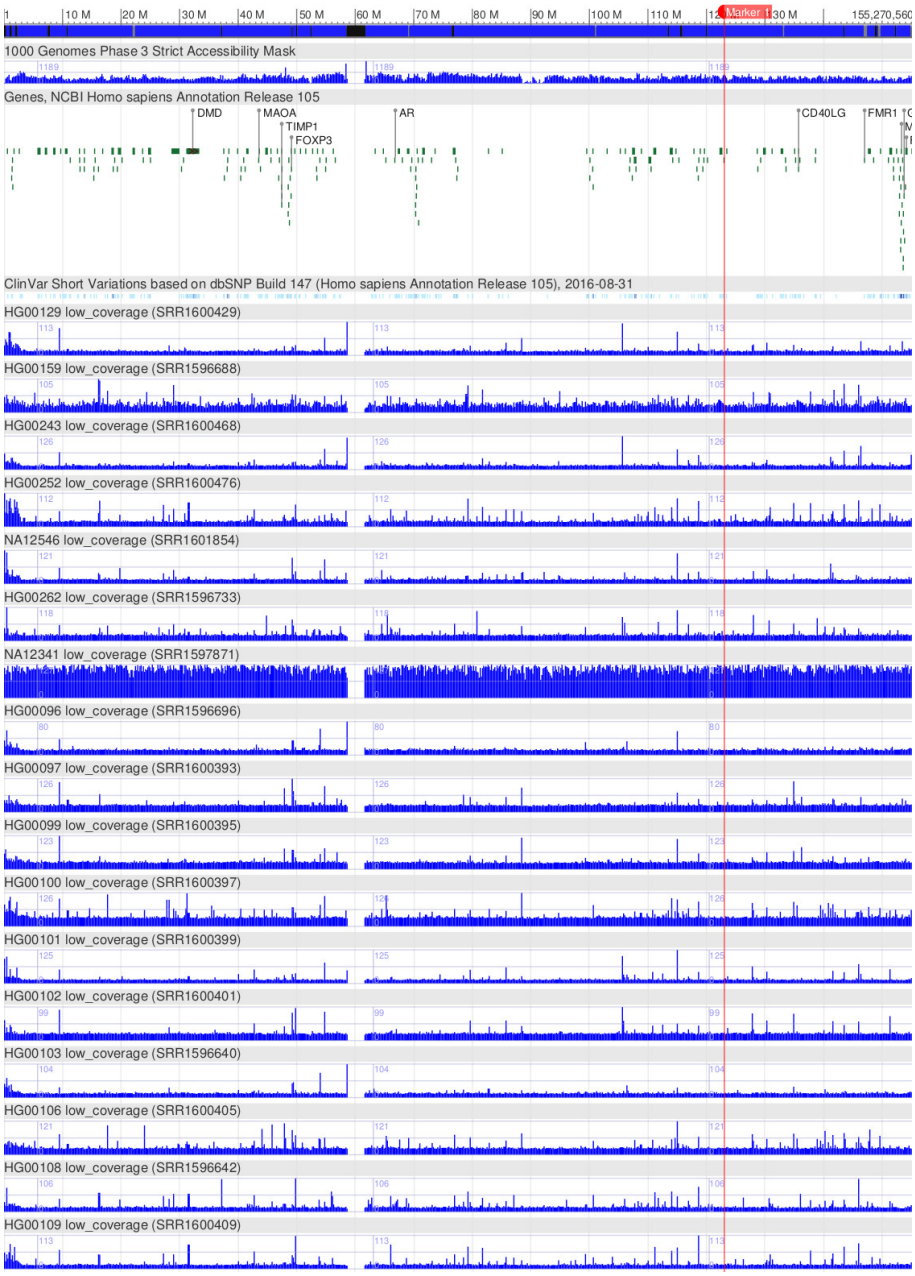


Figure I.15: Sequence coverage of the X chromosome for the 7 lymphoblastoid cell lines identified as having an over expression of XIAP (top 7 samples) and a subset of normal samples for comparison. There is no difference in the copy number around the XIAP gene (marked as red) compared to the rest of the chromosome. Nor is there a difference between the abnormal and the normal samples. This indicates that the over-expression is not due to a structural variation around the gene.

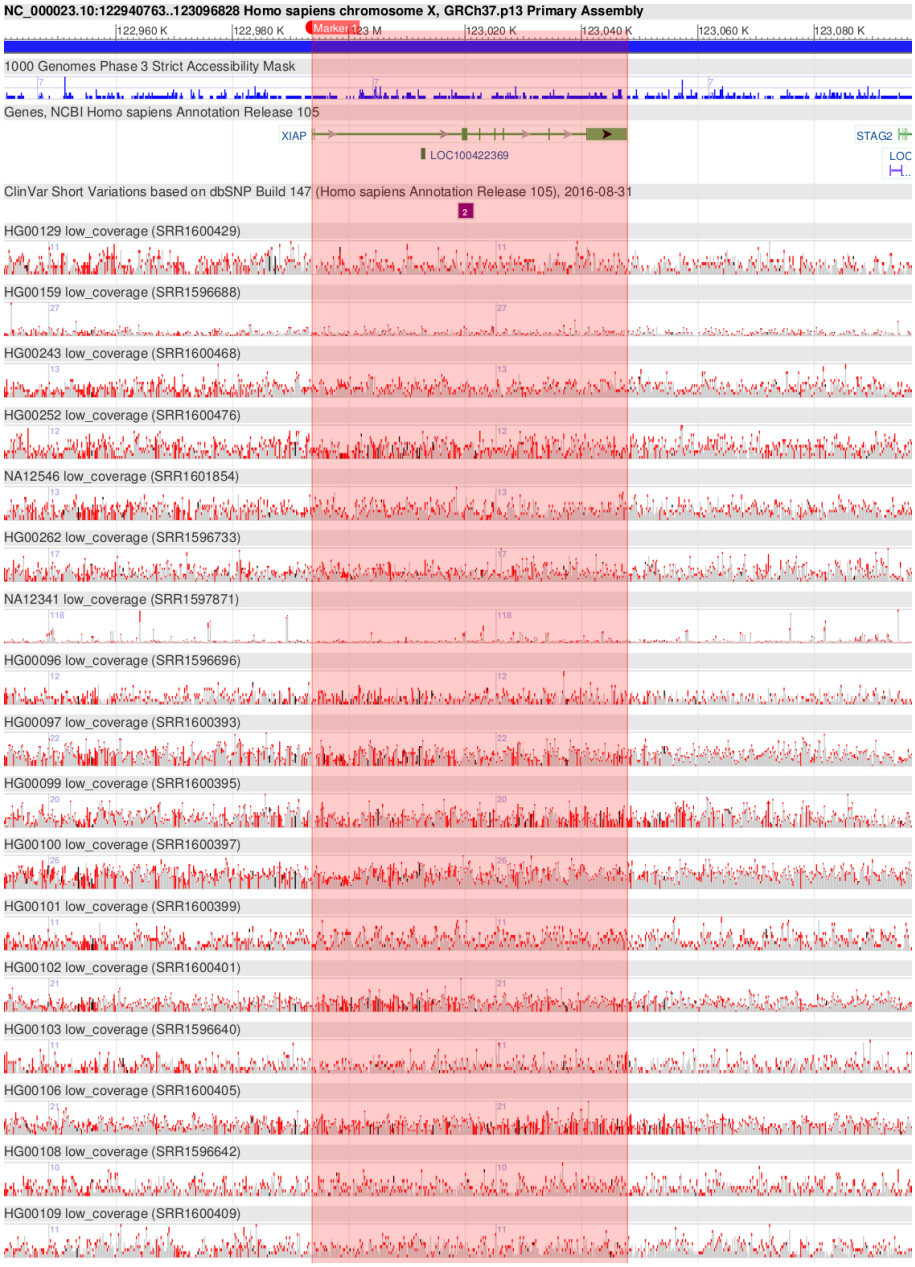


Figure I.16: Sequence coverage around the XIAP gene for the 7 lymphoblastoid cell lines identified as having an over-expression of XIAP (top 7 samples) and a subset of normal samples for comparison. There is no difference between the abnormal and normal samples, indicating that the over-expression is not due to a structural variation around the gene.

Bibliography

[1] Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Piñeros M, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. International Journal of Cancer. 2018 dec;144(8):ijc.31937.

[2] Wilson BE, Jacob S, Yap ML, Ferlay J, Bray F, Barton MB. Estimates of global chemotherapy demands and corresponding physician workforce requirements for 2018 and 2040: a population-based study. The Lancet Oncology. 2019 jun;20(6):769–780.

[3] Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell. 2000;100(1):57–70.

[4] Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011 mar;144(5):646–674.

[5] Weinberg RA. The biology of cancer. 2nd ed. New York: Garland Publishing Inc; 2014.

[6] Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, et al. Mutations of the BRAF gene in human cancer. Nature. 2002 jun;417(6892):949–954.

[7] Samuels Y, Wang Z, Bardelli A, Silliman N, Ptak J, Szabo S, et al. High Frequency of Mutations of the PIK3CA Gene in Human Cancers. Science. 2004 apr;304(5670):554–554.

[8] Yuan TL, Cantley LC. PI3K pathway alterations in cancer: variations on a theme. Oncogene. 2008 sep;27(41):5497–5510.

[9] Weinberg RA. pRb and Control of the Cell Cycle Clock. In: The biology of cancer. 2nd ed. New York: Garland Publishing Inc; 2014. p. 275–329.

[10] Sherr CJ, McCormick F. The RB and p53 pathways in cancer. Cancer Cell. 2002;2(AUGUST):103–112.

[11] Adams JM, Cory S. The Bcl-2 apoptotic switch in cancer development and therapy. Oncogene. 2007 feb;26(9):1324–1337.

[12] Strasser A, Harris AW, Huang DCS, Krammer PH, Cory S. Bcl-2 and Fas / APO-1 regulate distinct pathways to lymphocyte apoptosis. The EMBO journal. 1995;14(24):6136–6147.

[13] Huang DCS, Hahne M, Schroeter M, Frei K, Fontana A, Villunger A, et al. Activation of Fas by FasL induces apoptosis by a mechanism that cannot be blocked by Bcl-2 or Bcl-xL. Proceedings of the National Academy of Sciences. 1999 dec;96(26):14871–14876.

[14] Degenhardt K, Chen G, Lindsten T, White E. BAX and BAK mediate p53-independent suppression of tumorigenesis. Cancer Cell. 2002;2(September):193–203.

[15] George ML, Tutton MG, Janssen F, Arnaout A, Abulafi aM, Eccles Sa, et al. VEGF-A, VEGF-C, and VEGF-D in Colorectal Cancer Progression. Neoplasia. 2001;3(5):420–427.

[16] Valtola R, Salven P, Heikkilä P, Taipale J, Joensuu H, Rehn M, et al. VEGFR-3 and Its Ligand VEGF-C Are Associated with Angiogenesis in Breast Cancer. The American Journal of Pathology. 1999 may;154(5):1381–1390.

[17] Tsujii M, Kawano S, Tsuji S, Sawaoka H, Hori M, Dubois RN. Cyclooxygenase Regulates Angiogenesis Induced by Colon Cancer Cells. Cell. 1998;93:705–716.

[18] Rahden BHAV, Stein HJ, Pu F, Koch I, Langer R, Piontek G, et al. Coexpression of Cyclooxygenases (COX-1 , COX-2) and Vascular Endothelial Growth Factors (VEGF-A , VEGF-C) in Esophageal Adenocarcinoma. Cancer research. 2005;65(12):5038–5045.

[19] Skobe M, Hawighorst T, Jackson DG, Prevo R, Janes L, Velasco P, et al. Induction of tumor lymphangiogenesis by VEGF-C promotes breast cancer metastasis. Nature Medicine. 2001;7(2):192–198.

[20] Perl AK, Wilgenbus P, Dahl U, Semb H, Christofori G. A causal role for E-cadherin in the transition from adenoma to carcinoma. Nature. 1998;392(March):711–714.

[21] Vleminckx K, Vakaet L, Mareel M, Fiers W, Van Roy F. Genetic manipulation of E-cadherin expression by epithelial tumor cells reveals an invasion suppressor role. Cell. 1991 jul;66(1):107–119.

[22] Hazan RB, Phillips GR, Qiao RF, Norton L, Aaronson SA. Exogenous Expression of N-Cadherin in Breast Cancer Cells Induces Cell Migration , Invasion , and Metastasis. Journal of Cell Biology. 2000;148(4):779–790.

[23] Nieman MT, Prudoff RS, Johnson KR, Wheelock MJ. N-Cadherin Promotes Motility in Human Breast Cancer Cells Regardless of their E-Cadherin Expression. Journal of Cell Biology. 1999;147(3):631–643.

[24] Li G, Satyamoorthy K, Herlyn M. N-Cadherin-mediated Intercellular Interactions Promote Survival and Migration of Melanoma Cells. Cancer research. 2001;61:3819–3825.

[25] Pece S, Chiariello M, Murga C, Gutkind JS. Activation of the Protein Kinase Akt/PKB by the Formation of E-cadherin-mediated Cell-Cell Junctions. Journal of Biological Chemistry. 1999 jul;274(27):19347–19351.

[26] Miki Y, Swensen J, Shattuck-eidens D, Futreal PA, Tavtigian S, Liu Q, et al. A Strong Candidate for the Breast and Ovarian Cancer Susceptibility Gene BRCA1. Science. 1994;266(5182):66–71.

[27] Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, et al. Identification of the breast cancer susceptibility gene BRCA2. Nature. 1995 dec;378(6559):789–792.

[28] Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. Nature Genetics. 2008 jan;40(1):17–22.

[29] Nature, Glossary [Internet]: SNP [cited 2019 Jul 3];. Available from: `http://www.nature.com/scitable/definition/single-nucleotide-polymorphism-snp-295`.

[30] Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nature Genetics. 2013 apr;45(4):353–361.

[31] Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson Sa, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptorpositive breast cancer. Nature Genetics. 2007 jul;39(7):865–869.

[32] Schumacher FR, Al Olama AA, Berndt SI, Benlloch S, Ahmed M, Saunders EJ, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nature Genetics. 2018;50(7):928–936.

[33] Lin WY, Camp NJ, Ghoussaini M, Beesley J, Michailidou K, Hopper JL, et al. Identification and characterization of novel associations in the CASP8/ALS2CR12 region on chromosome 2 with breast cancer risk. Human Molecular Genetics. 2015 jan;24(1):285–298.

[34] Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature Genetics. 2007 jul;39(7):870–874.

[35] Ghoussaini M, Edwards SL, Michailidou K, Nord S, Cowper-Sal·lari R, Desai K, et al. Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. Nature Communications. 2014 dec;5(1):4999.

[36] Easton DF, Pooley Ka, Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature. 2007 jun;447(7148):1087–1093.

[37] Dunning AM, Healey CS, Baynes C, Maia AT, Scollen S, Vega A, et al. Association of ESR1 gene tagging SNPs with breast cancer risk. Human Molecular Genetics. 2009 mar;18(6):1131–1139.

[38] Bojesen SE, Pooley Ka, Johnatty SE, Beesley J, Michailidou K, Tyrer JP, et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. Nature Genetics. 2013 apr;45(4):371–384.

[39] Cox A, Dunning AM, Garcia-Closas M, Balasubramanian S, Reed MWR, Pooley Ka, et al. A common coding variant in CASP8 is associated with breast cancer risk. Nature Genetics. 2007 mar;39(3):352–358.

[40] Kruglyak L, Nickerson DA. Variation is the spice of life. Nature Genetics. 2001;27(march):234–236.

[41] Hinds Da, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, et al. Whole-genome patterns of common DNA variation in three human populations. Science. 2005 feb;307(5712):1072–1079.

[42] Meyer KB, O'Reilly M, Michailidou K, Carlebur S, Edwards SL, French JD, et al. Fine-Scale Mapping of the FGFR2 Breast Cancer Risk Locus: Putative Functional Variants Differentially Bind FOXA1 and E2F1. The American Journal of Human Genetics. 2013 dec;93(6):1046–1060.

[43] Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature Genetics. 2013 oct;45(10):1238–1243.

[44] Li Q, Seo JH, Stranger B, McKenna A, Pe'er I, LaFramboise T, et al. Integrative eQTL-Based Analyses Reveal the Biology of Breast Cancer Risk Loci. Cell. 2013 jan;152(3):633–641.

[45] Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017 nov;551(7678):92–94.

[46] Lee D, Lee GK, Yoon KA, Lee JS. Pathway-based analysis using genome-wide association data from a Korean non-small cell lung cancer study. PLoS ONE. 2013 jan;8(6):e65396.

[47] Fletcher MNC, Castro Maa, Wang X, de Santiago I, O'Reilly M, Chin SF, et al. Master regulators of FGFR2 signalling and breast cancer risk. Nature Communications. 2013 dec;4(1):2464.

[48] Quan B, Qi X, Yu Z, Jiang Y, Liao M, Wang G, et al. Pathway analysis of genome-wide association study and transcriptome data highlights new biological pathways in colorectal cancer. Molecular Genetics and Genomics. 2015 apr;290(2):603–610.

[49] Thomas R. Boolean formalization of genetic control circuits. Journal of Theoretical Biology. 1973 dec;42(3):563–585.

[50] Glass L, Kauffman SA. The Logical Analysis of Continuous , Non-linear Biochemical Control Networks. Journal of Theoretical Biology. 1973;39:103–129.

[51] Albert R, Othmer HG. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster. Journal of Theoretical Biology. 2003 jul;223(1):1–18.

[52] Dassow GV, Meir E, Munro EM, Odell GM. The segment polarity network is a robust developmental module. Nature. 2000;406(July):188–192.

[53] Fumiã HF, Martins ML. Boolean Network Model for Cancer Pathways: Predicting Carcinogenesis and Targeted Therapy Outcomes. PLoS ONE. 2013 jul;8(7):e69008.

[54] Noble WS. A Quick Guide to Organizing Computational Biology Projects. PLoS Computational Biology. 2009 jul;5(7):e1000424.

[55] Csikász-Nagy A, Battogtokh D, Chen KC, Novák B, Tyson JJ. Analysis of a Generic Model of Eukaryotic Cell-Cycle Regulation. Biophysical Journal. 2006 jun;90(12):4361–4379.

[56] Gérard C, Tyson JJ, Novák B. Minimal Models for Cell-Cycle Control Based on Competitive Inhibition and Multisite Phosphorylations of Cdk Substrates. Biophysical Journal. 2013 mar;104(6):1367–1379.

[57] Weis MC, Avva J, Jacobberger JW, Sreenath SN. A Data-Driven, Mathematical Model of Mammalian Cell Cycle Regulation. PLoS ONE. 2014 may;9(5):e97130.

[58] Novák B, Tyson JJ. A model for restriction point control of the mammalian cell cycle. Journal of theoretical biology. 2004 oct;230(4):563–79.

[59] Jain HV, Nör JE, Jackson TL. Modeling the VEGFBcl-2CXCL8 Pathway in Intratumoral Agiogenesis. Bulletin of Mathematical Biology. 2008 jan;70(1):89–117.

[60] Jain HV, Nör JE, Jackson TL. Quantification of endothelial cell-targeted anti-Bcl-2 therapy and its suppression of tumor growth and vascularization. Molecular cancer therapeutics. 2009 oct;8(10):2926–36.

[61] Nguyen LK, Cavadas MaS, Scholz CC, Fitzpatrick SF, Bruning U, Cummins EP, et al. A dynamic model of the hypoxia-inducible factor 1 (HIF-1) network. Journal of Cell Science. 2013 mar;126(6):1454–1463.

[62] Kim M, Reed D, Rejniak Ka. The formation of tight tumor clusters affects the efficacy of cell cycle inhibitors: A hybrid model study. Journal of Theoretical Biology. 2014 jul;352:31–50.

[63] Wang Z, Zhang L, Sagotsky J, Deisboeck TS. Simulating non-small cell lung cancer with a multiscale agent-based model. Theoretical Biology and Medical Modelling. 2007;4(1):50.

[64] Olsen MM, Siegelmann HT. Multiscale Agent-based Model of Tumor Angiogenesis. Procedia Computer Science. 2013;18:1016–1025.

[65] Gerlee P, Anderson ARA. A hybrid cellular automaton model of clonal evolution in cancer: The emergence of the glycolytic phenotype. Journal of Theoretical Biology. 2008 feb;250(4):705–722.

[66] Araujo A, Baum B, Bentley P. The Role of Chromosome Missegregation in Cancer Development: A Theoretical Approach Using Agent-Based Modelling. PLoS ONE. 2013 aug;8(8):e72206.

[67] Anderson ARA, Weaver AM, Cummings PT, Quaranta V. Tumor Morphology and Phenotypic Evolution Driven by Selective Pressure from the Microenvironment. Cell. 2006 dec;127(5):905–915.

[68] Caravagna G, Giarratano Y, Ramazzotti D, Tomlinson I, Graham TA, Sanguinetti G, et al. Detecting repeated cancer evolution from multi-region tumor sequencing data. Nature Methods. 2018;15(9):707–714.

[69] Diaz-Uriarte R. Cancer progression models and fitness landscapes: A many-to-many relationship. Bioinformatics. 2018;34(5):836–844.

[70] Hosseini SR, DIaz-Uriarte R, Markowetz F, Beerenwinkel N. Estimating the predictability of cancer evolution. Bioinformatics. 2019;35(14):i389–i397.

[71] Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. Nature Genetics. 2016;48(3):238–244.

[72] Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, et al. A big bang model of human colorectal tumor growth. Nature Genetics. 2015;47(3):209–216.

[73] GWAS Catalog [Internet]; European Bioinformatics Institute; [cited 2018 Dec 1];. Available from: `https://www.ebi.ac.uk/gwas/efotraits/EFO_0000311`.

[74] Göring HHH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole Sa, et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nature Genetics. 2007 oct;39(10):1208–1216.

[75] Consortium TGO. Gene ontologie: Tool for the unification of biology. Nature Genetics. 2000;25(1):25–29.

[76] Carbon S, Dietze H, Lewis SE, Mungall CJ, Munoz-Torres MC, Basu S, et al. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Research. 2017 jan;45(D1):D331–D338.

[77] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000 jan;28(1):27–30.

[78] Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research. 2016 jan;44(D1):D457–D462.

[79] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research. 2017 jan;45(D1):D353–D361.

[80] Lloyd CM, Lawson JR, Hunter PJ, Nielsen PF. The CellML Model Repository. Bioinformatics. 2008 sep;24(18):2122–2123.

[81] Chelliah V, Juty N, Ajmera I, Ali R, Dumousseau M, Glont M, et al. BioModels: ten-year anniversary. Nucleic Acids Research. 2015 jan;43(D1):D542–D548.

[82] Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Research. 2012 nov;41(D1):D808–D815.

[83] Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research. 2014 jan;42(D1):D1001–D1006.

[84] Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Research. 2012 jan;40(D1):D930–D934.

[85] Blood eQTL browser [Internet]: Groningen, Gene Network; [cited 2019 Jul 3];. Available from: `https://genenetwork.nl/bloodeqtlbrowser/`.

[86] STRING v9.1 [Internet]; EMBL; [cited 2019 Jul 3];. Available from: `http://string91.embl.de/`.

[87] Faure A, Naldi A, Chaouiya C, Thieffry D. Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. Bioinformatics. 2006 jul;22(14):e124–e131.

[88] Irons DJ. Logical analysis of the budding yeast cell cycle. Journal of Theoretical Biology. 2009 apr;257(4):543–559.

[89] Warmerdam DO, Kanaar R. Dealing with DNA damage: Relationships between checkpoint and repair pathways. Mutation Research/Reviews in Mutation Research. 2010 apr;704(1-3):2–11.

[90] Shaltiel IA, Krenning L, Bruinsma W, Medema RH. The same, only different - DNA damage checkpoints and their reversal throughout the cell cycle. Journal of Cell Science. 2015 feb;128(4):607–620.

[91] Schlatter R, Schmich K, Lutz A, Trefzger J, Sawodny O, Ederer M, et al. Modeling the TNFα-Induced Apoptosis Pathway in Hepatocytes. PLoS ONE. 2011 apr;6(4):e18646.

[92] Brancho D, Ventura Jj, Jaeschke A, Doran B, Flavell Ra, Davis RJ. Role of MLK3 in the Regulation of Mitogen-Activated Protein Kinase Signaling Cascades. Molecular and Cellular Biology. 2005;25(9):3670–3681.

[93] Tu CC, Kumar VB, Day CH, Kuo WW, Yeh SP, Chen RJ, et al. Estrogen receptor α (ESR1) over-expression mediated apoptosis in Hep3B cells by binding with SP1 proteins. Journal of Molecular Endocrinology. 2013 aug;51(1):203–212.

[94] Li Xb, Jiao S, Sun H, Xue J, Zhao Wt, Fan L, et al. The orphan nuclear receptor EAR2 is overexpressed in colorectal cancer and it regulates survivability of colon cancer cells. Cancer Letters. 2011 oct;309(2):137–144.

[95] Eissing T, Conzelmann H, Gilles ED, Allgöwer F, Bullinger E, Scheurich P. Bistability Analyses of a Caspase Activation Model for Receptor-induced Apoptosis. Journal of Biological Chemistry. 2004 aug;279(35):36892–36897.

[96] Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, et al. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proceedings of the National Academy of Sciences. 2011 nov;108(44):18026–18031.

[97] Parisien M, Khoury S, Chabot-Doré AJ, Sotocinal SG, Slade GD, Smith SB, et al. Effect of Human Genetic Variability on Gene Expression in Dorsal Root Ganglia and Association with Pain Phenotypes. Cell Reports. 2017 may;19(9):1940–1952.

[98] Morris MD. Factorial Sampling Plans for Preliminary Computational Experiments. Technometrics. 1991;33(2):161–174.

[99] Sobol IM. Sensitivity Estimates for Nonlinear Mathematical Models. Mathematical Modeling and Computational Experiment. 1993;1(4):407–414.

[100] Sobol IM. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation. 2001 feb;55(1-3):271–280.

[101] Rand DA. Mapping global sensitivity of cellular network dynamics: sensitivity heat maps and a global summation law. Journal of The Royal Society Interface. 2008 aug;5:S59–S69.

[102] Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP. Universally Sloppy Parameter Sensitivities in Systems Biology Models. PLoS Computational Biology. 2007;3(10):e189.

[103] Cavoretto R, De Rossi A, Perracchione E, Venturino E. Graphical Representation of Separatrices of Attraction Basins in Two and Three-Dimensional Dynamical Systems. International Journal of Computational Methods. 2017 feb;14(01):1750008.

[104] Cavoretto R, Rossi AD, Perracchione E, Venturino E. Reliable approximation of separatrix manifolds in competition models with safety niches. International Journal of Computer Mathematics. 2015 sep;92(9):1826–1837.

[105] McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 nov;491(7422):56–65.

[106] Lappalainen T, Sammeth M, Friedländer MR, t Hoen PAC, Monlong J, Rivas Ma, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013 sep;501(7468):506–511.

[107] GDC Data Portal [Internet]; NIH [updated 2019 Jun 5, cited 2019 Jul 3];. Available from: `https://portal.gdc.cancer.gov/`.

[108] GDC Data Portal: Legacy archive [Internet]; NIH [cited 2019 Jul 3];. Available from: `https://portal.gdc.cancer.gov/legacy-archive/search/f`.

[109] GenomeWideSNP_6.hg19.bed [Internet]; Thermo Fisher Scientific Inc [updated: 2016 Jun 2, cited 2019 Jul 3];. Available from: `http://www.affymetrix.com/Auth/analysis/downloads/lf/genotyping/GenomeWideSNP_6/GenomeWideSNP_6.hg19.bed.zip`

[110] GenomeWideSNP_6.na35.annot.csv [Internet]; Thermo Fisher Scientific Inc [updated: 2015 Apr 30, cited: 2019 Jul 3];. Available from:

http://www.affymetrix.com/Auth/analysis/downloads/na35/genotyping/GenomeWideSNP_6.na35.annot.csv.zip

[111] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007 sep;81(3):559–575.

[112] Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015 oct;526(7571):68–74.

[113] Shapiro SS, Wilk MB. An Analysis of Variance Test for Normality (Complete Samples). Biometrika. 1965;52(3/4):591–611.

[114] Gu L, Zhu N, Zhang H, Durden DL, Feng Y, Zhou M. Regulation of XIAP Translation and Induction by MDM2 following Irradiation. Cancer Cell. 2009 may;15(5):363–375.

[115] AlQarni S, Al-Sheikh Y, Campbell D, Drotar M, Hannigan A, Boyle S, et al. Lymphomas driven by EpsteinBarr virus nuclear antigen-1 (EBNA1) are dependant upon Mdm2. Oncogene. 2018 jul;37(29):3998–4012.

[116] Ferreira MA, Gamazon ER, Al-Ejeh F, Aittomäki K, Andrulis IL, Anton-Culver H, et al. Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. Nature Communications. 2019 dec;10(1):1741.