Drug target identification for potential diseases using mathematical models

Arturo Blázquez Navarro

Theoretical Biophysics

Humboldt Universitat zu Berlin

A thesis submitted for the degree of

MSc Biotechnology

July 2012

Abstract

Der Einsatz von Methoden der Systembiologie bei der Suche nach potentiellen Targets ist ein vielversprechender Ansatz, um biologische Vorkenntnisse für die Entwicklung von effektiven Medikamenten zu verbessern. Obwohl noch in den Anfängen steckend, haben diese Methoden bereits bewiesen, dass sie in diesem Prozess der Medikamentenentwicklung nutzlich sein und die Entscheidungsfindung der Target-Auswahl verbessern konnen. In meiner Arbeit werden Ansatze der Systembiologie verwendet, um eine Methode zur Target-Auswahl über biologische Modelle von Gewöhnlichen Differentialgleichungen (ODE) zu entwickeln.

Es existieren bereits verschiedene auf ODE-Modellen basierende Ansatze zur Target-Entdeckung. Zum Beispiel bietet TIde (50) einen weitreichenden Rahmen für die Target-Vorhersage auf Modellen, die Pathologien beschreiben. Es bietet eine Reihe von möglichen Modifikatoren, um den gesunden Zustand eines biochemischen Netzwerkes wiederherzustellen, sowie die Targets auf die diese Modifikationen wirken. Allerdings hat diese Herangehensweise einige inhärente Beschränkungen, da die meisten relevanten ODE Modelle keine Krankheiten beschreiben, sondern normale physiologische Zustande. Weiterhin konnen komplexe Erkrankungen, wie beispielsweise Krebs, verschiedene Ursachen haben, die nicht individuell modelliert werden konnen. Daher ist das Ziel meiner Arbeit ein Tool zu erstellen, welches an einem gesunden Modell Veranderungen simuliert (Mutationen, usw.) sowie potentielle Medikamente findet, die den gesunden Phanotyp wiederherstellen. Diese Herangehensweise liefert einen Rahmen zur Target-Vorhersage basierend auf "gesunde" Modellen. Die grundlegende Idee meiner Arbeit ist es, ausgehend von einem Model zur Beschreibung eines gesunden Organismus eine Reihe von modifizierten Modellen zu erstellen, um mögliche Mutationen im System zu simulieren. Dies wird durch die systematische Beeintrachtigung jeder Reaktion und jeder Spezies im Modell realisiert. Der mutierte Modellsatz wird dann auf die Abweichung von den Charakteristika bewertet, die

besonders relevant für die Funktion des Systems sind. Danach werden die pathologischen Systeme im Rahmen von TIde auf Behandlungen gescannt, die den gesunden Zustand wiederherstellen. Zuletzt wird ein Satz von Modellen mit den verschiedenen modellierten Mutationen ausgegeben sowie alle Behandlungen, die den ursprunglichen gesunden Zustand wiederherstellen.

Um die Ergebnisse des Tools zu analysieren, wurde die Methode auf zwei verschiedene Modelle des TGF-β Signalweges angewandt worden, die an vielen menschlichen Krankheiten, darunter Krebs, beteiligt ist. Die Ergebnisse stimmen mit den biologischen Erkenntnissen uberein und bieten genugend Informationen, um potenzielle Targets in einem gesunden System vorzuschlagen.

The incorporation of Systems Biology methods into target selection is a promising approach to improve prior biological knowledge towards a better drug development. Although still at their infancy, these methods have already proved to be useful in the process of drug development and could therefore further improve decision making in target selection. In my work, Systems Biology approaches have been used to develop a method for target selection based on biological models described by Ordinary Differential Equations (ODEs).

Various approaches do already exist for drug target discovery based on ODE models. For example, TIde (50) offers a useful framework for drug-target prediction on models describing pathologies, providing a list of possible modifiers that can restore the biochemical network to a healthy state and the targets of these modifiers. However, this kind of approach has some inherent limitations, as most relevant ODE models do not describe diseases, but the normal physiological state. Furthermore, many complex diseases, such as cancer, can have many different causes, which cannot be modelled individually. Therefore, the main objective of my work is creating a tool that simulates the possible alterations (mutations, etc.) that could modify a healthy model, and to find possible drugs that can restore the healthy phenotype, thus creating a framework for drug-target prediction based on "healthy" models. The general idea behind my work is, starting from a model describing a healthy organism, to create a set of modified models in order to simulate possible mutations in the system. This is done by systematically impairing every reaction and species in the model. The mutated model set is then evaluated for the deviation from the characteristics considered most relevant for the functioning of the system. After this, the pathological systems are scanned for treatments that can restore the healthy state using the framework provided by TIde. The final output is a set of models, with different modelled mutations and all the treatments that were able to restore the original healthy conditions.

In order to analyse the results provided by the tool, the method has been applied on two different models of the TGF-β signalling route, which is involved in many human diseases including cancer. The obtained results are consistent with biological evidence and provide enough information to hypothesise potential drug targets in a healthy system.

Acknowledgements

First, I would like to thank my supervisor Prof. Dr. Dr. h. c. Edda Klipp, for proposing this research topic to me, creating the main idea behind my methodological approach and for her continuous support during my stay, providing a highly supportive scientific environment at her group.

This work would not have been possible without Marvin Schulz, who not only created TIde, the tool on which my work is based, provided many helpful ideas for the methodological part and assisted me with the writing of this thesis, but also helped me throughout this whole time, responding to my (continuous) doubts and teaching me everything I had to know for my research topic. Thanks a lot, Marvin.

I would also like to thank Dr. Zhike Zi, for his useful comments on the results provided by my methodology and Dr. Katja Wengler, for providing me access to her TGF-β model.

It is a great pleasure to thank my colleagues at the Theoretical Biophysics group, to Dr. Matteo Barberis, Wolfgang Giese, Friedemann Uschner, Falko Krause, Margit Heiske, Matthias Reis, Dr. Orianne Mazemondet and Dr. Samuel Drulhe, and especially my "comrades-in-arms" at the Student Room, Lydia Farack, Matthias Bock, Andr Dautcourt, l Vrtesy and Stephan Adler, for their support and readiness to help in all the problems I had.

I am truly thankful to Prof. Dr. Lynne Yenush for her help as project tutor at my home university, and for recommending this fantastic Theoretical Biophysics group as the right place to do my Master Thesis.

I would also like to show my gratitude to Sabine Wagnitz, for her continuous assistance during this complex journey through the bureaucracy that implies doing your Master Thesis as an Erasmus exchange student. I don't think I would have survived it without your help.

I also want to acknowledge funding from the Sonderforschungsbereich (SFB) 618.

And last, but not least, I want to thank all those that have been at my side during this year, my family, my partner and all my friends. You listened to me when I was desperate because nothing worked the way it should and encouraged me to forget about difficulties and do my best. Thanks a lot.

Contents

List of Figures

List of Tables

1

Introduction

1.1 Drug Discovery

Drug development is a complex process involving biotechnology, molecular biology, pharmacology and medicine. It is roughly composed of six steps. The first step is the election of an adequate system for the evaluation of drug candidates, which can be one or more proteins, genes, other biomolecules, tissue cultures or a living organism. This election is crucial to the success of the whole process, since it will be used as a reference for assessing the effectivity of drug candidates. It therefore has to simulate the *in vivo* effects of the drug in an acceptable way and avoid, as well as possible, the rejection of potentially useful drug candidates and the selection of toxic or ineffective drug candidates. The second step is screening and selecting drug candidates for their behaviour in the selected system. Large libraries of chemicals are screened and selectivity and/or efficacy of the compounds are evaluated. Selected drug candidates can also be examined using cross-screening, for evaluating the activity of the drug on other systems and predicting toxicity. Afterwards, drug candidates are optimised by using rational structural criteria, to increase activity and selectivity, reduce toxicity and improve the drug kinetics (ADME) of the drug in the living organism.

The fourth step are pre-clinical trials. Drug candidates are evaluated *in vitro* and *in vivo* for their safety, undergoing pharmacodynamics and pharmacokinetics assays and to determine adequate doses for trials on humans. If the drug candidate has an adequate performance in pre-clinical trials, it goes through clinical trials. Clinical trials are composed of three phases: Phase I, usually performed on healthy individuals (except for antitumoural drugs) to determine drug safety; Phase II, performed on small groups of patients, to preliminarily determine effectivity

and safety on patients; and Phase III, performed on big groups of patients to finally determine effectivity and safety of the drug. A drug candidate that performs acceptably in clinical trials, can then be submitted for approval. If the drug is approved by the relevant authority, it is then commercialised. Drug effects are monitored post-commercialisation for effectivity and safety, in what is also called Phase IV.

Developing a drug is a long and expensive process. It lasts for around ten years and with costs in the order of 1 billion dollars (10^9\$) (1). Each further step in drug development supposes an exponential growth of costs. A failure on the late clinical assays therefore means a high economical damage for the company developing the drug. Because of this, a greater emphasis has been put on refining the performance of the first steps, which are comparably much cheaper and less time-consuming with respect to the whole process. (2)

1.1.1 Approaches to Drug Evaluation

The first step of drug development and probably the most crucial one for the success of the process, is the election of an adequate system for the evaluation of the drug candidates. There are basically two approaches to this election: pathology-centred and target-centred. Pathology-centred approaches are based on biological models for the pathological condition, which can be either complete organisms or tissue cultures.Alternatively, target-centred approaches are based on drug targets, biological molecules thought to be relevant for the illness, whose behaviour is intended to be changed through the use of drugs. (3)

1.1.1.1 Pathology-centred approaches

Pathology-centred approaches have been the preferred approaches for drug discovery for decades, resulting in the discovery of thousands of drugs, which are still successfully used. These approaches regard the biological model as a black box, not considering the mechanisms behind pathology and drug action, but the physiological effect of the drug candidates in the considered model. (3)

For this, a relevant biological model has to be chosen i.e. a bacteria for developing a new antibiotic, tissue cultures or yeast for well-characterised illnesses or whole animals for evaluating complex diseases. For massive screening of compounds, tissue cultures and unicellular organisms are preferred over whole animals, although whole animals might be better predictors of drug effects on humans. Even though this approach does in principle not require understanding the aetiology, it does however rely on the assumption that relevant molecular mechanisms in

the model are similar to that of the illness in human. This assumption can be incorrect, as similar phenotypes can be caused by different mechanisms. Therefore, in most cases, at least a superficial knowledge of the aetiology is needed for the choice of a biological model. However, as it does not require a prior knowledge of the molecular mechanisms causing the illness, this approach is therefore less susceptible to errors stemming from inexactitudes on understanding of the illness. (3)

Due to their complexity, biological models can be good predictors for drug effects. They can take into account, not only the efficacy and effectivity of the drug, but also the side effects, toxicity and even the pharmacokinetics of the drug. Moreover, they can predict the effects of multi-target drugs, off-target effects, drugs with unknown mechanisms, etc. Because of all this, results obtained by pathology-centred approaches can be more easily translated into a commercial drug. (3)

However, these approaches are not without disadvantages. They depend absolutely on the existence of a relevant biological model, so they cannot be used for diseases for which there is no such a model, which can be the case for complex diseases, such as psychiatric diseases. Another disadvantage is the difficulty for the optimisation of a candidate drug without prior knowledge of the mechanism of action, since no structural information is known of the drug target(s) and therefore an improvement on the performance of the drug can only be done, either by random modifications or using general rules for optimisation. (3)

As stated above, the choice of an adequate model is not trivial and it depends on many variables, such as the pathology that is being studied, prior knowledge of good models for the pathology and economical and ethical issues. Therefore, a balance between costs and performance benefits of each possible biological model have to be taken into account. Nevertheless, results obtained using these models are not necessarily reproducible in humans, even if the model appears to be a good predictor of drug effects. (3)

1.1.1.2 Target-centred approaches

Target-centred approaches regard the living organism as a series of biomolecules, genes and pathways, some of them involved in the development of the studied pathology. The activity of these components is altered as a result of illness, so that the objective is to modify their activities back to normal, restoring healthy conditions. Alternatively, the objective can be to defeat the robustness of a living organism, such as a pathogen or a tumoural cell. In either case, one or more *drug targets* are chosen. Drug targets are biomolecules which are believed to be key in

the functioning of one or more pathways and are therefore capable of modulating their activity. Drug candidates are evaluated for their ability to modify the activity of these drug targets. (4)

Targets are biomolecules showing with a key role in disease development, so that modulating its action provokes a significant therapeutic response. Targets can be either have a causative role in the disease or the ability to modify or alleviate the disease process or its symptoms. Possible targets can be determined by identifying molecules that are upregulated or downregulated in disease conditions by using transcriptomics or proteomics assays, and further biological knowledge; or alternatively by developing a mechanistic model on the causes of the disease, based on all the information available. In the first case, it cannot be assured that the selected target is causing the disease and is not just a consequence of it. The second approach, although it incorporates knowledge from different sources and not only correlation values, it is not without problems, as knowledge on disease development is still very limited. (4)

The main advantage of target-centred approaches is that they permit drug development for diseases which cannot be easily studied on physiological models. Moreover, they facilitate the optimisation of drug candidates, as these approaches are based on prior knowledge about the illnesses, permitting rational design and modification of the drug candidates. However, because of this dependence on prior knowledge, results derived from these analyses can be of difficult or no application to developing a therapy, as they can be based on erroneous data or, more importantly, on partial data. To avoid this, targets must be validated before drug screening. But target validation is very complex and depends upon the availability of valid predictive disease models, requiring a demonstration that modulating the target behaviour permits a substantial modification in disease in such a model. (4)

Target-centred approaches represent inherently a simplification of the drug action: one or more biomolecules are used as the model for evaluating the a drug, predicting its action for systems composed by thousands of biomolecules. It therefore does not take into account interactions with most of the biomolecules, so that cross-screening assays must be performed to predict possible drug toxicity or side effects and automatically disregards other possible mechanisms to influence on the biomolecule, different to targeting such as transcription or pre- and post-translational modulation. (4)

1.1.2 Attrition in Drug Development

Drug target-centred approaches substituted about 20 years ago pathology-centred approaches as the main focus for drug research and design of new drug candidates, basing drug discovery

on rational criteria and knowledge of the molecular mechanisms behind pathology. At the time these approaches were introduced it was believed that their use would increase drug productivity, compensating the deficit in pharmaceutical innovation predicted for the first decade of 2000 (4, 5). However, the effect seems to have been the opposite (6).

The pharmaceutical industry is currently facing an unprecedented challenge for its business model. Confidence of consumers has fallen, health-care budgets have diminished and key patents are expiring. But the key problem is the decay in R& D productivity: although there has been an increase on investment, the number of approved innovative drugs has not risen accordingly, but remained stable for the last decade or even reduced, while most discovered drugs are based on well-established targets (6). Roughly one in four compounds fails after submission, incurring the full development costs but not conveying any economical benefit (2), attrition rates are very high (7), while the average cost of developing a drug is rising rapidly. This has led some experts to state that if there is no dramatic increase in R& D productivity, we may be moving to the extinction of pharmaceutical industry as we understand it today, with fatal consequences for human health (6).

Although many factors which could affect R&D productivity have changed in the past 10 years (6) it has been observed that this decrease on efficiency has coincided in time with the introduction of target-based drug discovery. Therefore, it has been hypothesised that there is a causal relation between these two events (3, 4). Among the arguments for this relation, is the striking observation that, while target-based approaches have been the focus of drug-discovery approaches, only a minority of the approved innovative drugs correspond to the result of drug-target centred approaches, whereas the discoveries corresponding to pathology-centred approaches are still majority. This could be due to the main limitations of target-centred approach, as it is inherently reductionist and depends totally on our (limited) knowledge on pathological processes. However, despite the evidence that target-centred approaches are not proportioning the expected results, there are reasons to think that this is not a problem of the approach*per se*, but of the way it is applied. Namely, how targets are chosen and validated. (4)

Target-centred approaches have proven to be very effective for developing novel treatments for validated targets (for example, the target of an already existing drug) (3, 4). Therefore, various improvements have been suggested to respond the question of how to improve target selection, with the objective of maximising efficacy and minimising toxicity, the two main causes for drug attrition (2). These improvements basically rely on emphasising a more thorough previous study of the selected targets and on early validation of the target. Firstly, all available

information relevant to the disease, the selected target and the relations between them has to be gathered. This must include, not only the biological and biochemical information, but also pharmaceutical and economical information, such as alternative treatments or feasibility of clinical testing. Especially, one must exclude the possibility that pathological alterations in the drug target are a consequence and not a cause of the disease, as many of the studies used for target selection examine correlation, but not causativity. Secondly, there must be a greater emphasis on early target validation. When a drug against the target is not available, other approaches, such as interference RNA technology, can be used to examine target validity (4). These changes would increase costs and duration of the early phases of drug development. However, as inappropriate target selection (6) and selection of the drug candidates regardless of the mechanism of drug-target interaction (3) have been cited as possible reasons behind low productivity, this increase of spend on early phases could reduce total costs of drug developments (as less drugs would be rejected at later stages), increase efficiency and therefore increase R&D productivity.

The incorporation of Systems Biology methods to target selection is a promising approach to improve prior biological knowledge towards a better target selection. Although still on their infancy, these methods have already been proved useful in the process of drug development and could therefore be used to improve decision making in target selection.

1.2 Systems Biology

The sequencing of the human genome by Human Genome Project has opened an unprecedented development on many aspects of biology. The so-called *omic* technologies, comprising among others genomics, transcriptomics, proteomics or metabolomics, provide researchers with a previously unknown quantity of data on many aspects of biology, such as genetics, gene transcription and translation and their regulation, metabolic and signalling pathways, etc. The results obtained by these high throughput techniques are, in most cases, too complex to be interpreted by an unaided human. Thus, there has been an effort for developing bioinformatical tools, to classify and interpret the data obtained by the use of these technologies. As a result of the complexity of the handled data, it is in many cases not possible to formulate an *a priori* hypothesis of what the behaviour of the system will be and therefore design an experiment to confirm or reject the hypothesis; but rather, experiments are designed without a previous hypothesis, to generate a great quantity of data, which will be analysed and used to generate hypotheses.

Therefore, there has been a shift on research orientation, from a reductionist approach to discovery (hypothesis-driven science), where researchers focus in very concrete topics, a biomolecule, a gene, etc. to a maximalist discovery-driven science, genome-wide data centred. This shift has finally led to the appearance of a new paradigm in biology which aims to close the gap between hypothesis-driven science and discovery-driven science. This new paradigm is known as systems biology.(8)

Systems biology is a novel approach to understanding biology in a holistic and integrated way. It considers living beings as complex systems (and not as the sum of its components), where relations between components are as important as the components themselves. The study of one or more key individual components is abandoned in favour of focusing in the dynamic behaviour of systems as a whole. Systems studied by this approach can comprise from the most general level (ecosystems) to the most specific (biochemical pathways), each of them a compendium of all levels itself, as these systems are not closed or unrelated but interact and interchange continuously information with each other.(8)

The objective of systems biology is creating descriptions of these systems (models) that not only account for their known behaviour, but also can accurately predict their behaviour under additional conditions, before they have been experimentally assayed. Hypotheses generated by the model can be verified or falsified using directed experiments, thus contributing to an improvement of the model's accuracy. This cycle of hypothesis generation and model refinement, where models are iteratively improved in accordance to new experimental information has been termed "cycle of systems biology" accordingly. (9)

This cycle is considered the fundamental core of systems biology and is the basis for the integration between hypothesis-driven science and discovery-driven science. Descriptions of the systems are generally based on data from high-throughput technology, but once the model has been built, it is used for hypothesis generation, allowing small-scale research. The confirmation or rejection of the hypotheses leads, in turn, to a refining of the model accordingly. The model therefore comprises an integration of all available information, both on the macro and micro level. (9)

This is, in definitive, the objective of systems biology: relating individual and systemic properties, revealing the properties of complex biological systems and the laws behind them with the ultimate objective of uncovering the fundamental principles behind life.(9)

1.2.1 Computational Models in Systems Biology

A biological computational model is an abstract description of a biological process which represents an integration of the available data in an objective and unambiguous way. Models can be formulated using different mathematical frameworks, depending on the biological characteristics of the network. They are modular, formed by single small entities (proteins, genes, reactions, modifications, etc.) and, as such, highly flexible and modifiable. They can be used to simulate different conditions and perturbations in the given system and can be visually represented in an easily understandable way. Because of all this, computational models are nowadays the basis of systems biology understanding. They describe our knowledge on biological systems, their components and interactions, as a whole. However, they are not only a description of the available knowledge, but they can be useful to extend our knowledge: we can use them as a predictive tool leading to testable hypotheses. (9)

1.2.1.1 Types of Computational Models

Biological systems can be modelled using different mathematical approaches, depending on the biological system, the nature of the data and the subject of our research. These approaches can be classified, for increasing levels of complexity of the solutions:

- **Statistical mining models**. Statistical mining models are based purely based on numerical correlations, including no mechanistic data. These models differentiate between dependent and independent components, reducing the biological behaviour to a principal components problem, thus elucidating how changes in certain variables affect dependent variables. Statistical mining is the most basic level of biological modelling and can be used for physiological-level modelling. (8)

- **Boolean networks**. Boolean networks are systems of Boolean variables, i.e. variables which can only adopt two states (true or false). In these models, states are deterministically determined by the states of other variables in the network. Time is a discrete variable and on each unit of time, the state values of the network are recalculated from the previous state. These models are often used to simulate gene regulation. In these models, genes can be either on (transcriptionally active) or off (9) (transcriptionally inactive) and depending on the activity of other genes, they can become on or off after a number of time steps.

- **Discrete models**. Discrete models are an extension of boolean models, allowing for variables to have a finite number of states, making it possible to describe systems where components can adopt more than two different states. (9)

- **Ordinary differential equation (ODE) models**. In ODE models, variables and time are continuous. The values of the variables are defined by ODEs, which describe the variation of the values over time. While the derivative of a variable can be easily calculated at any time point, the state can only be calculated by integration, which in many cases is too complex to be solved analytically, making numerical approximations necessary. Knowing the initial concentration of the species and the kinetics of the reactions they participate in, a biochemical reaction network can be explicitly modelled using ODE. Because of this, ODE models can be useful for modelling metabolic and signalling networks. (9)

- **Partial differential equation models**. Partial differential equations are differential equations that incorporate multivariable functions. They can model processes that are dependent on more than one variables, such as time and one or more spatial dimensions. Because of that, they can be appropriate to model non-homogeneous biochemical processes, where concentrations of a species varies within a compartment. (9)

- **Bayesian models**. Bayesian models are probability models based on conditioned (Bayesian) probability. They are composed of a set of random variables, where relations between consequences and their probable causes are established. Therefore, causes are modelled as probabilistically dependent on their consequences. This makes them useful for modelling biological systems where the consequences are more easily observed than the causes, such as gene expression or cell signalling. (9)

- **Stochastic models**. Stochastic models are models that take into account random effects and their influence on system behaviour. In these models, the result of reactions is modelled following a probability distribution and are therefore, nondeterministic. Because of this, stochastic approaches can be useful for modelling biological systems with high random influences, such as those where the number of interacting molecules is very low, so interactions between them depend on random encounters. (9)

In this work, we focus on ODE models for the search of possible drug targets in biochemical networks, because they are a balance between simplicity and quality of the results, as the simplest modelling framework that provides continuous quantitative data.

1.2.1.2 Development of Computational Models

The development of a model is not trivial. Many decisions that will affect its usefulness and potential applications have to be taken in advance, as models can be developed with different focuses, different levels of detail and different mathematical approaches. In the end, the question is how to design a model, so that it not require the use of non-identifiable parameters (parameters which cannot be determined using available data), but still is able to describe the relevant system in an accurate way and opens the possibility to predictions of model behaviour.

Developing a biological model usually begins with a compilation of the available data: underlying mechanisms, available experimental results and other models describing the same or similar systems. These data can be used to build a primary model which will be used as a basis for further development. As knowledge in genomics, transcriptomics and proteomics has developed, the increasing amount of data has made collecting this information a very laborious work, since it is generally dispersed in literature. To aid in the building of primary models, a number of databases have been developed e.g. Gene Ontology (10), KEGG (11), Brenda (12) or Reactome (13), which centralise relevant information, such as relations between biological concepts, interactions between molecules, data on reaction kinetics, etc. If there already exist computational models describing the same pathway or similar pathways, it can be interesting to use them as a start point for the development of the new model, taking advantage of the knowledge they comprise, adding new information and correcting their main weaknesses. Computational models are easily accessible in repositories, such as BioModels (14) and JWS online (15). Tools like semanticSBML (16) can be extremely helpful for the retrieval and fusion of models.

From all the compiled data, a primary or scaffold model is built. Models are written using the Systems Biology Markup Language (SBML) (17), which describes models as a set of species in a compartment, which interact through reactions. Scaffold models can be built using various SBML compatible tools such as CellDesigner (18), a visual model editor and COPASI (19), a more mathematical-centred editor, which also permits simulation of model behaviour.

After building the primary model, it has to be fitted to the existing experimental results. In many cases, not all the parameters of the system have been determined as they cannot be found on the literature. Therefore, these parameters have to be estimated. This is a critical step for the elaboration of the model. The number of parameters that has to be estimated deeply affects the quality of the resulting model, as some of the parameters can be unidentifiable, i.e. they are indistinguishable from other parameters and cannot be accurately estimated. The probability

of having unidentifiable parameters increases with the number of parameters that need to be estimated and decreases with increasing complexity of the network. The model undergoes a series of optimisation steps, until it fits acceptably the existing experimental data. The fitted model is then ready for hypothesis generation. (9)

1.2.2 Systems Biology and Drug Discovery

Drug discovery has traditionally been a systems-centred approach. Before the advent of scientific medicine, remedies were discovered by observation of their effects on diseased and healthy people, the most complex and relevant model to drug discovery. After the late 1800s and early 1900s, there was a modernisation of drug development process, where new compounds were obtained by chemical synthesis and the use of human models substituted in favour of animal models, therefore maintaining the systemic approach. Indeed, most of the drugs that we use today, such as aspirin or Prozac, still derive from the use of this traditional *systems* approach. (20)

As explained above, coinciding with the rise in popularity of target-centred approaches in drug discovery, there has been a worrying decrease in pharmaceutical R&D productivity. This has been attributed to deficiencies in target identification or validation. Systems biology approaches offer a promising alternative for drug target identification, drug *in silico* testing or pharmacokinetics and secondary effects simulation. Computational tools are comparatively cheap and permit the obtention of abundant data. The use of this technology could thus contribute to a decrease in costs and an increase of efficiency in these fields.

1.2.2.1 Modelling of Physiological and Disease Processes

There have been numerous efforts to modelling disease relevant processes. These efforts vary in the scale (organism, tissue or cell), in the type of modelled process, level of detail or mathematical framework and can be useful for simulating different aspects of pathology study. Computational models have been used to simulate the relationship between disease and certain physiological or genetic markers, tissue and cell level mechanisms causing disease, pathology-relevant biochemical pathways, etc. For the creation of models specifically aimed towards drug target simulation, it is necessary to incorporate a sufficient level of detail, such that the possible drug targets are modelled accurately. For example, to evaluate for possible drug targets in a signalling route, it is necessary that all the proteins in the route are explicitly modelled and not as a *black box*, for the model to be a useful predictor of drug targets. (20)

Computational models in general and signalling pathways in particular, have already given some promising results, such as the ability to accurately predict the results of a clinical assay (21). However, the use of computational models for drug target evaluation is still at its very first steps. Successes of model predictions could be considered anecdotal, as computational models have not been optimised for drug target evaluation, which is only possible with clinical data. Therefore, models are not yet at a stage where decisions taking can be relied upon them, but will remain for the next years only one of many inputs for decision making. (20)

1.2.2.2 Modelling of Signalling Pathways

Signalling pathways are biochemical pathways that transmit information content within a cell. They inform the cell about the extracellular physicochemical conditions, including nutrients, hormones and toxins and modulate the cell response towards these stimuli. They are generally well conserved through evolution, consisting of a membrane receptor which activates an effector when binding a ligand, this effector triggers a cascade of protein interactions, which in the end regulate many aspects of cell physiology, including gene expression.

Signalling pathways have been related to many complex diseases, including cancer. Therefore, a great effort has been put on modelling the main signalling pathways. Some key human signalling pathways have already been extensively studied, such as the Wnt (22), Jak/STAT (23), NFκB (24), TGF-β (25) and EGFR/ErbB (26). However, the use of signalling pathways models for drug target prediction is at a very early phase, even when compared to other models for disease. This is due to particular problems of model developing for signalling pathways. The main limitation that has to be faced is the availability of data of sufficient quality to enable modelling and qualifying the pathways structure and dynamics. Despite the development of high-throughput techniques, the obtained data is generally not of adequate type and quality for model development. However, there are still some encouraging results, as, for instance the identification of ErbB3 as a possible therapeutic target of the ErbB route (27), which show the potential of signalling pathways modelling for drug target evaluations. (28)

1.3 TGF-β Signalling

Transforming growth factor beta (TGF-β) isoforms are a group of polypeptide factors with an important role in different aspects of cell homeostasis regulation in many vertebrates, including humans (29). Among them, TGF-β 1 is considered the prototypical member of the family.

This family contains over 35 structurally related factors which includes other TGF-β isoforms (2 and 3) as well as related cytokines. TGF-β 1 is a secreted homodimeric protein that binds specifically to membrane receptors, triggering a signalling response that regulates numerous cellular responses, such as proliferation, differentiation, migration and apoptosis. Malfunctioning in the signalling pathway has been found to be implicated in several human diseases, including cancer (30).

Because of the importance of TGF-β signalling in human disease, there has been a great interest in studying the pathway using systems biology approaches. In fact, several mathematical models have already been published that describe the TGF-β signalling pathway (25, 31, 32, 33, 34) with different levels of detail and focus on different aspects of TGF-β signalling.

1.3.1 TGF-β Signalling Pathway

In the canonical TGF-β signalling pathway, TGF-β isoform 1 (TGF-β1) binds to the two receptor molecules, type I (TβRI) and type II (TβRII). These receptors are transmembrane proteins, with an intracellular Ser-Thr kinase domain. Binding of the ligand induces the formation of heterotetrameric TβRI-TβRI-TβRII-TβRII complex and the phosphorylation of TβRI GS-rich-domain by the constitutively active TβRII. This enables the GS-domain-mediated recruitment of the receptor-regulated Smads (R-Smads), Smad2 and Smad3 and their phosphorylation by TβRI on two Ser residues at their C termini. Phosphorylated R-Smads form heteromeric complexes with Smad4, the common mediator Smad, as well as homomeric complexes, of unknown function. The heteromeric Smad4-R-Smad complexes function as secondary messengers of the TGF-β signalling pathway and accumulate in the nucleus, where they regulate transcription of target genes, both positively and negatively. (30, 35)

However, TGF-β signalling is not a linear, unregulated pathway. A third type of Smad proteins, inhibitory Smads (I-Smads), may play an important role at pathway regulation, as they can trigger signal termination. I-Smads can antagonise TGF-β signalling by distinct mechanisms. They are competitive inhibitors of TβRI-TGF-β binding and inhibit interactions between R-Smads and Smad4. They can recruit Smurf1 and 2, which induce TβRI ubiquitination and degradation. Furthermore, they can directly repress Smad-induced transcriptional responses (36). Since some of the components of these inhibitory pathways, such as Smad7 (an I-Smad) (37) and Ski/SnoN (38), are upregulated by TGF-β signalling, they may be part of a negative feedback mechanism. Nevertheless, the role of I-Smads in actual physiological conditions is still rather controversial (35). Furthermore, there exist other alternative TGF-β signalling routes

apart from the canonical TGF-β pathway. These pathways can be derived by the TGF-β receptors directly interacting with other non-Smad proteins, thus initiating a parallel signalling. These routes can also modulate the activity of the canonical pathway, interact with other pathways and transmit the TGF-β signal directly to other pathways (cross-talking) (39).

1.3.2 TGF-β Signalling and Disease

As a consequence of the central role of TGF-β signalling in key processes, including the cell cycle, this signalling pathway is connected to many human diseases. Malfunctioning of the TGF-β signalling pathway has been connected to a number of human diseases, such as Marfan syndrome (40), fibrotic conditions and malignancies (29): TGF-β shows an increased activity in a number of fibrotic disorders, causing an accumulation of connective tissue in lung, kidney, liver or other organs. Abnormal activity of this pathway has been linked to inflammatory disorders as well. Studies on mutant mice have revealed that alterations in TGF-β signalling pathway have deep effects on the development or homeostasis of many organs, therefore showing that alterations in signalling can be a cause for illness(29).

In cancer, TGFβ has a complicated role. During early stages of tumourogenesis, it acts as a cancer suppressor, inducing apoptosis and inhibiting cell proliferation. Because of this, genetic inactivation of TGF-β signalling through Smads has been found to be inactivated in human cancer. For example, about 50% of pancreatic carcinoma have been shown to have a loss of Smad4 expression, while in colorectal cancer there can be a loss of a single copy of TβRI (TGF-β receptor I). Moreover, many oncoproteins have been found to directly interact with or post-translationally modify Smads, interfering with TGF-β signalling. (29)

However, at later stages of cancer development, TGF-β shows the opposite effect. Tumour cells usually show an increased production of TGF-β, promoting metastasis, angiogenesis and a suppression of the immune response. This is known as the TGF-β switch in cancer progression, which triggers epithelial-mesenchymal transition (EMT). EMT is a process through which epithelial cells generate transitory mesenchymal derivatives with increased motility that colonise new sites. This process, vital in embryo development, is activated during cancer progression by many different factors, including TGF-β, which regulate transcription factors activated during embryogenesis. (29)

Targeting TGF-β signalling has been tackled as a possible therapy against certain tumours. As a result of the complicated role of TGF-β in cancer, both approaches to cancer suppression, enhancing and hindering signalling have been successfully used in preclinical and clinical assays:

upregulation of Arkadia, a complex that enhances TGF-β signalling, provides tumour suppression in colorectal cancer in murine models (41), while PSK, a protein-bound polysaccharide which strongly inhibits TGF-β through an unclear mechanism, has shown beneficial therapeutic effects in clinical assays for various types of tumours and is used as a treatment for gastric and colorectal cancers in Japan (42). These two findings highlight the importance and complexity of TGF-β signalling in tumour physiology.

These characteristics, together with the availability of numerous computational models, make TGF-β signalling a very attractive target for the systems biology approaches.

2

Aim of the project

The aim of this project is creating a semi-automatic framework for drug target identification in computational models describing physiological processes. Starting from an ODE model that describes the normal behaviour of biochemical network, the objective is to simulate potential pathological conditions and propose different combinations of drugs that can restore the initial behaviour of the model. The application of this tool leads to a systematical selection of those network components that could be interesting drug targets against these alterations.

3

Methods

3.1 ODE modelling

ODE models are mathematical descriptions of processes quantified by their change in time based on an ordinary differential equation (ODE) system. They are continuous, uniform and deterministic. Because of that, they can be good predictors only when the number of molecules in the system is large enough, so that these three assumptions are realistic. A large number of molecules implies that the quantity can be modelled as a continuous variable, that molecules can diffuse instantaneously within their compartments and that random effects do not exert a notorious influence in their behaviour. Most molecules in biological systems exist in very high numbers, making ODE modelling suitable for many cases. However, other molecules, especially transcription factors and signal transducers, tend to be in much lower numbers. The concentration of these molecules can be on the order of only 10 molecules per cell (43). These systems are generally subject to other approaches, such as stochastic modelling. (44)

3.1.1 Structure of biochemical ODE models

The basic elements of an ODE model are a set of variables and their change in time. In a biological model, variables correspond to the existing species, with the numerical value corresponding to their concentration, amount or activity. These variables are modified by a set of processes, comprising biochemical reactions and transport mechanisms, which change the concentration or activity values of the species. Variations on the variables values are modelled using a system of ODEs. (44)

An ODE is an equation containing a function of one variable and its derivatives, depending on one independent variable x. An ODE can generally adopt two types of forms:

Implicit ODE:

$$F(x, y, y', ..., y^{(n)}) = 0$$

Explicit ODE:

$$y^{(n)} = f(x, y, y', ..., y^{(n-1)})$$

In biological ODE models, equations generally adopt the explicit form: the dependent variable is defined as the concentration or activity of a species, while the independent variable is time t. For example, for a simple chemical reaction

$$S \rightarrow 2P \tag{3.1}$$

the change in time of the concentrations can be described using the ODE system:

$$\begin{aligned} \frac{dS}{dt} &= -v \\ \frac{dP}{dt} &= 2v. \end{aligned} \tag{3.2}$$

The ODEs describe that the species P is produced at a rate which is twice of consumption rate of S. Knowing the initial value of the variables S_0 and P_0, the concentration of the species can be determined for any time t by solving the equations analytically.

For a biochemical network consisting of m species and r reactions, its behaviour as a system can be generally described using the following system of equations, where n_{ij} is the stoichiometric coefficient (number of molecules produced or consumed by a single reaction event) of the species i in the reaction j and v_j is the velocity of reaction j:

$$\frac{dS_i}{dt} = \sum_{j=1}^{r} n_{ij} v_j \text{ for } i = 1, ..., m.$$

With this information and the initial values for the species, the concentration or activity of each species can be calculated at any time point solving the ODE system. Only very simple ODE systems, such as (3.2), can be solved analytically. As systems are generally more complex, the solution for the system has to be computed using numerical approximations. (9, 44)

3.1.2 Simulations of ODE models

The most basic approach for the numerical approximation of an ODE is the Euler method. For a variable y dependent on time t, this approach approximates the value of the $y(t+h)$ to

$$y(t+k) \approx y(t) + h \cdot y'(t),$$

where $y(t)$ is a known value of y and h is the step size, the difference between the estimated time point and the reference one. Since y' is defined by the ODE system for every time point, we can calculate $y(t)$ by step-wise iterating

$$y_{n+1} \approx y_n + h \cdot y'_n$$

for the whole parameter space. This method provides a way of calculating the value of any variable for any value of t . But only approximately, because unlike analytical methods, numerical methods introduce an error in calculations, which in this case is proportional to h^2. Therefore, the smaller that the step size h is, the more accurate results are. (9)

Euler's method is not generally used as an ODE solver, but it is an interesting method to show the general way how solvers work. It is also the basis for the construction of other methods. Numerical methods used for ODE solving are nowadays more complex and exact. Examples of solvers currently used in ODE modelling are approaches based on Runge-Kutta method, LSODA (45) and CVODE (comprised in SUNDIALS package) (46).

3.1.3 Stoichiometric matrix

The stoichiometric coefficients determine the relations between species within a biochemical network. They influence not only the individual properties of each reaction, but also affect the global structural properties of the whole network. For example, in a simple reaction, such as (3.1), the stoichiometric constants determine the production rate of P ($\frac{dP}{dt}$) as double of the consumption rate of S ($-\frac{dS}{dt}$). In more complex systems, the stoichiometric constants can determine global properties, such as the existence of a non-trivial steady-state. (9, 44)

When dealing with such complex systems, it can be very useful to express the stoichiometric coefficients of every species in every reaction of the network as a matrix, in what is called the stoichiometric matrix. For example, for the network

$$\overset{v_1}{\rightleftharpoons} \quad S_1 \quad \overset{v_2}{\rightarrow} \quad 5S_2$$
$$v_3 \uparrow\downarrow$$
$$2S_3$$

the stoichiometric matrix is

$$N = \begin{pmatrix} 1 & -1 & -1 \\ 0 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix},$$

where the columns describe the stoichiometries of every reaction and rows that of every species. In order to determine the signs, the convention for reversible reactions is assigning as positive reactions "from left to right" and "from top to down". For irreversible reactions, the sign is assigned accordingly with the direction of the reaction. (44)

The stoichiometric matrix contains information on the structure of the network, such as combinations of fluxes that can result in a steady state, dead-ends, ramifications or relations between species. Moreover, it is useful for handling big biochemical networks, since it permits separating the structure of the model, determined by the matrix, from the kinetic information. (9, 44)

3.1.4 Reaction kinetics

Mass-action kinetics are the most widely used reaction kinetics for this work. They are based on the Law of Mass Action, discovered by Guldberg and Waage in the 19th century(47). This law states that the reaction rate is proportional to the probability of collision of the reactants, which is in turn proportional to their concentration powered to their absolute stoichiometric constant. For example, for the simple reaction

$$S_1 + 2S_2 \rightarrow P,$$

the reaction rate is

$$v = kS_1S_2^2,$$

where k is the proportionality constant known as kinetic constant. (44)

In the case that the reaction is reversible, both forward and backwards reactions are taken into account, so that the reaction rate of

$$S_1 + 2S_2 \rightleftharpoons P$$

is defined as

$$v = k_+S_1S_2^2 - k_-P.$$

This can be expressed described using the ODE system

$$\frac{dS_1}{dt} = -k_+ S_1 S_2^2 - k_- P$$
$$\frac{dS_2}{dt} = -k_+ S_1 S_2^2 - k_- P$$
$$\frac{dP}{dt} = k_- P - k_+ S_1 S_2^2.$$

The general mass action kinetics for a reaction with substrate concentrations S_i and product P_j is

$$v = k_+ \prod_i S_i^{|m_i|} - k_- \prod_j P_j^{|m_j|},$$

where m is the stoichiometric constant of the reactant and $k_- = 0$ for irreversible reactions.

Mass action kinetics are good predictors of the behaviour for non-catalysed reactions. This makes it a very useful approach for modelling, for example interactions between proteins. But for enzymatic reactions, which are of great importance in biology, other kinetic models have to be used.

The most widely used model for enzyme kinetics is Michaelis-Menten. In Michaelis-Menten, the mechanism of an enzymatic reaction is modelled as

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_2}{\to} E + P, \tag{3.3}$$

where E is an enzyme, S is its substrate and P the product. Therefore, using mass action kinetics, the system can be described by the equations

$$\begin{aligned} \frac{dS}{dt} &= -k_1 E \cdot S + k_{-1} ES \\ \frac{dP}{dt} &= k_2 ES \\ \frac{dE}{dt} &= (k_{-1} + k_2) ES - k_1 E \cdot S \\ \frac{dES}{dt} &= k_1 E \cdot S - (k_{-1} + k_2) ES, \end{aligned} \tag{3.4}$$

Since this system cannot be solved analytically, a few assumptions are taken to simplify it. If $k_1, k_{-1} >> k_2$ and $S(t = 0) >> E$, we observe that a quasi steady state is reached for the enzymatic complex, so that

$$\frac{dES}{dt} = k_1 E \cdot S - (k_{-1} + k_2) ES = 0.$$

This can be reordered as

$$ES = \frac{E_{total}S}{S + \frac{k_{-1}+k_2}{k_1}},$$

where $E_{total} = E + ES$. Since $v = \frac{dP}{dt} = k_2ES$ and ES is maximal if all enzyme is in complex ($E_{total} = ES$), the maximal velocity of the reaction can be described as $v_{max} = k_2E_{total}$. The reaction rate yields the following equation:

$$k_2ES = \frac{k_2E_{total}S}{S + \frac{k_{-1}+k_2}{k_1}} \rightarrow v = \frac{v_{max}S}{S + \frac{k_{-1}+k_2}{k_1}}$$

which can be simplified to

$$v = \frac{v_{max}S}{S + K_m},$$

where $K_m = \frac{k_{-1}+k_2}{k_1} \approx \frac{k_{-1}}{k_1}$.

The change of S (3.3) can then be described as:

$$\frac{dS}{dt} = -\frac{v_{max}S}{S + K_m}$$

$$\frac{dP}{dt} = \frac{v_{max}S}{S + K_m}.$$

Michaelis-Menten equation thus permits a simplification of the whole ODE system shown in (3.4) to a system that is only dependent on one variable, substrate S and two constants, maximal velocity (v_{max}) and the Michaelis constant (K_m), which is equivalent to the quantity of substrate that fulfils $v = \frac{v_{max}}{2}$. This equation can be adapted for reversible reactions and, interestingly for drug target search, different kinds of enzyme inhibition can be easily modelled in the equation as changes in the constants v_{max} and K_m, depending on the inhibition mechanism. (44)

Apart from mass action and Michaelis-Menten kinetics, other kinetic laws can be found in ODE models. For example, a reaction can be described as a constant flux or as dependent on non-reactant species, showing essential and non-essential activations or repressions effects. A variety of different kinetic functions are included in modelling software for the description of other complex behaviours, including allosterism, cooperativity, substrate or product inhibitions, etc.

3.2 Parameter estimation

The mathematical description of a biochemical network consists of four types of information which can be defined as vectors or matrices: the vector $S(t)$ which describes the concentration values at any time point, the stoichiometric matrix N, the vector v, which describes the kinetic equations that determine the shape of the functions of S versus time and the parameter vector p, which contributes to the numerical value of v. The first three components are often well-known, while a part of the values of p usually has to be determined. This parameter vector comprises concentrations of external species affecting the system and kinetic parameters, such as proportionality, maximal velocity or Michaelis-Menten constants. In principle, although these parameters could be experimentally determined, in many cases this is not possible, as data may be too sparse or of the wrong type. Furthermore, parameter values may contain measurement errors and differences in measurement conditions have been found to lead to high variances in parameter values within the literature. Because of this, most parameters have to be estimated after the ODE model has been built. (9, 44)

This is done by fitting the model to the experimental data. If the model was adequately designed and experimental data were complete and exact, this procedure would allow for a determination of the true parameter set. This is not possible, however, as experimental data always contain a certain amount of noise. Therefore, computational methods are used to estimate approximate parameter values that can be used for biological modelling. (44)

The process of parameter estimation is based on two assumptions. Firstly, the experimental data must be deterministically produced by the processes described in the ODE model, with a set of unknown parameters. Secondly, experimental data contains some independent and normally distributed noise. Therefore, since a perfect fitting is not possible, as implied by the second assumption, parameter estimation is performed with the objective of obtaining a parameter set such that the experimental data are the most likely to be produced. (44)

3.2.1 Concept of objective function

Experimental data of a biological system can be expressed as a time course, a plot of one or more dependent variables of the system versus time. In a biological context, these variables can be the concentration of a species, the flux through a reaction, or a ratio of more than one variables. The time course is used as the reference to which the model is fitted during parameter

estimation. At the parameter estimation, the parameters are optimised so that it holds that

$$\overline{y}(t_p) \approx y(t_p) \text{ for all values of } p$$

where $\overline{y}$ is the experimental value determined by the time course, y is the value predicted by the model and t_p is the time for every measurement. (9)

This optimisation is based on the use of an objective function, which determines the quality of the model fit to the experimental data. The objective function is a measure of the deviation of the model from the experimental data. It depends on the time course data $\overline{y}$ and the model simulation data $y(t)$, for every variable i and time point p, and adopts the form of a sum of squared residuals:

$$f = \sum_{i=1}^{I} \sum_{p=1}^{P} \frac{(\overline{y}_i(t_p) - y_i(t_p))^2}{\sigma} \tag{3.5}$$

The value of the objective function is the reference for selecting between all possible parameter sets. This is done by performing an optimisation, scanning the parameter space (all possible combinations of parameter values) in the search of minima in the function that lead to a model fitting which is considered as satisfactory. (9)

3.2.2 Local and Global Optimisation

Model fitting leads to a optimisation problem where the objective function (3.5) is evaluated for local or global minima. A local minimum is a parameter set p that satisfies the condition that no other parameter set in the vicinity leads to a lower value of f. If there exists no other value in the parameter space for which f is lower, this point is called a global minimum. The scan for minima of f can be restricted to a region of the parameter space, by introducing constraints in the optimisation, such as defining intervals of allowed values for the parameters. (44)

For solving the optimisation problem, computational tools called optimisers are used. There exist two basic kinds of optimisers: local and global optimisers. Local optimisers search for local minima by evaluating the vicinity of a starting point for vector p and improving step by step the optimisation, until no value for p offers a lower value of the objective function. Global optimisers investigate the whole parameter space, in order to find the most optimal solutions. As this can be impossible due to the number of needed simulations, global optimisation is normally achieved by scanning the parameter space by a series of random jumps. As the search of global minima

is, in principle, not dependent on the starting point for the optimisation, this method tends to be more robust, while local optimisation requires a smaller number of simulations to achieve an acceptable solution, making them more efficient. Besides them, there exist other hybrid methods for objective function optimisation which combine the advantages of both methods. Popular optimisers for parameter estimation include BFGS (48), which is a local optimiser, and simulated annealing (49), which is a global optimiser. (9, 44)

3.2.3 Identifiability

In parameter estimation, different parameter sets can produce the same value of the objective function. This means that a single parameter set cannot be chosen as optimal and numerical optimisations might yield different parameter sets. This condition, where at least one parameter cannot be assigned to a single value, is known as non-identifiability. (44)

Non-identifiability can be of two different kinds: structural or practical. Structural non-identifiability stems from the mathematical structure of the model. This provokes that, for any possible experimental data, regardless of their quantity, type or their level of detail, more than one parameter sets can equally reproduce the same simulation results with respect to measurable variables. This can happen, for example, if two parameters a and b always appear in the model as the product $c = a \cdot b$. In this case, any combination of a and b with the same value of c will lead to the same result of the simulation. Structural non-identifiabilities can therefore be resolved by modifying the mathematical structure of the model.

Practical non-identifiabilities arise when, although the model may be theoretically identifiable, experimental data are not of sufficient quality to determine the approximate values of all parameters. This is more likely to happen if the model is overparameterised, i.e. if it contains a higher number of parameters than those needed to model the experimental results. (44)

3.2.4 Model Fitting and Prediction

During parameter estimation, the ODE model is fitted to the time course dataset, so that the resulting model is able to accurately reproduce the results for the same experimental conditions. This does not automatically make the model a valid tool for the prediction of novel results, and a model that fits perfectly the available data might be a very poor predictor under different conditions. This can be caused by an overfitting of the model. Overfitted models are models that reproduce the experimental results better than the *true* model, which is the model that theoretically determines the results. As a consequence of this, they will fit better the known

experimental data than those provided by further experiments. Overfitting can be caused a high amount of noise in data and insufficient number of data points so that random effects have a great influence on data distribution. The use of an unnecessarily high number of parameters (overparameterisation) can also favour the overfitting of the model. (44)

Since one of the objectives of ODE modelling is the generation of new predictions, overfitting should be avoided. Overfittings can be detected by cross-validation. Cross-validation is performed by dividing the experimental dataset in two subsets: one used for model fitting and the other for prediction testing. The model is optimised for the first data subset, while the second data subset is not used in model construction, but for assessing the accurateness of the predictions of the model.This procedure can be repeated for many datasets to evaluate the quality of predictions of the model. (44)

3.3 TIde: SBML based drug-target identification

TIde (50) is an open-source program for the systematic scanning of drug targets in ODE models which are implemented in the Systems Biology Markup Language (SBML) (17). TIde is available both as an online web-program at `http://lynx.biologie.hu-berlin.de/TIde/default/index` and as a part of the semanticSBML package, on `http://sourceforge.net/projects/semanticsbml/`

This program is a useful framework for the simulation of drug effects in ODE biochemical models and for drug target detection. This whole work is based on the framework provided by TIde, chosen as a reference for drug-target identification because of its flexibility, automatability and the simplicity of the approach. Therefore, the parts of the framework on which this work is based will be explained in detail.

3.3.1 Drug target identification as a parameter estimation problem

The general approach of TIde for drug target search consists in converting the evaluation of possible drug targets within an ODE model to a general parameter estimation problem. The program incorporates possible drugs to the network as new parameters and proceeds to the optimisation providing as the final output a list of potential drugs, their target and molecular mechanism. (51)

For the formulation of the problem, an accurate description of both the pathological and the healthy state of a biological system are needed. In this framework, the pathological state

is modelled explicitly as an ODE model. This model can describe, for example, the working metabolism of a pathogenic organism, biochemical processes in a cancer cell or abnormal activity of a signalling route in a disease. The healthy state that we want to achieve through drug targeting is described using a time course, i.e. concentrations of species at certain time points. For the case of cancer or pathogen organisms, this time course can describe alterations in the system's dynamic response leading to cell death. For other diseases, the time course can describe the normal behaviour of the system. To calculate the deviation between the pathological model and the healthy time course, an objective function is defined. This function is minimised when the pathological model fits the healthy data, i. e. when the behaviour of the model is considered to be that of a healthy system. (51)

Starting from the pathological model, TIde models a set of possible drugs by manipulating the kinetics of the model to include inhibitors targeting all reactions in the system, with various mechanisms of action. The concentration of the inhibitor set is stated as an unknown parameter, creating a vector of parameters p which has to be estimated. The resulting model is then fitted to the time course data, by evaluating the objective function for minima. The existence and complexity of acceptable drug combinations depends on many factors. These include the structure and dynamics of the biochemical network, the concentration and number of allowed drugs and the complexity of the objective function. Therefore, in order to solve the problem, different optimisation strategies can be followed, depending on the complexity of the problem. This optimisation finally results in a list of inhibitor concentrations that sufficiently satisfy the healthy conditions. (51)

3.3.2 Modelling of Theoretical Drugs

In the TIde framework, the parameters of the optimisation problem represent the action of potential drugs systematically targeting every possible drug target in the biochemical network. In most cases, these parameters are not included as a part of the model and therefore, the model has to be manipulated to include the new variables. This is done in the following way:

For every selected reaction, the kinetics are modified by including all possible modifiers (inhibitors or non-essential activators), with different modes for action, which depend on the type of kinetic. Kinetics are identified by numerical evaluation and modified employing an internal library, built based on the Systems Biology Ontology (SBO) (52). New kinetics of the

system are introduced as superimposed kinetic formulae of the form

$$v' = v\frac{1+\frac{A}{k_A}}{1+\frac{I}{k_I}}$$

for constant flux, mass action law and other kinetics, where v' and v are the new and original reaction rates, respectively, A is the concentration of non-essential activator, I the inhibitor concentration and k is the binding constant of the modifier. For Michaelis-Menten kinetics, further inhibitors with different modes of action are introduced as

$$v' = \frac{v_{max}S}{S + K_m\left(\frac{1+\frac{I_{comp}}{k_{comp}}}{1+\frac{I_{unc}}{k_{unc}}}\right)} \cdot \frac{1+\frac{A}{k_A}}{(1+\frac{I_{non}}{k_{non}})(1+\frac{I_{unc}}{k_{unc}})}$$

where I_{comp} is a competitive inhibitor, I_{non} is a non-competitive inhibitor and I_{unc} is an uncompetitive inhibitor. For all cases, the binding constant of the modifiers is equal to 1 ($k = 1$). (51)

These superimposed equations permit incorporating various possible modifications to the systems kinetics without affecting the structure, as the resulting ODE model is equivalent to the non-manipulated one when the concentration of all modifiers is 0 and to normal activation/inhibition kinetics when only one modifier in the system has a non-zero concentration. (51)

3.3.3 Description of the healthy state

In this framework, the ODE model is used to describe the pathological state of a biochemical system with an adequate level of detail and accurateness. The model is evaluated for a set of modifier concentrations which simulate a drug treatment, in order to restore a healthy state. These healthy state conditions are provided by the user as a set of concentration time points $\overline{y}_i(t_j)$ for different species i at different times j. The time course describes the molecular characteristics of the healthy state, including those species which are altered as a result of the disease and other important metabolites, to assess whether the simulated treatment shows side-effects. In the end, an objective function is constructed, describing the conditions that have to be fulfilled so that the original model can be considered to be in healthy state. (51)

For each time point $\overline{y}_i(t_j)$ of the time course, three possible conditions can be acceptable. If the substance i needs to be below a certain concentration $\overline{y}_i(t_j)$, a solution will be considered acceptable if $y_i(t_j, \theta) < \overline{y}(t_j)$, where $y_i(t_j, \theta)$ is the concentration of the species i in the model

y for a drug set θ. If the concentration of i has to be below a certain value, an acceptable solution will fulfil $y_i(t_j, \theta) > \overline{y}_i(t_j)$. Lastly, if the concentration of i has to be within an interval $(\overline{y}_i(t_j) - \sigma_i(t_j), \overline{y}_i(t_j) + \sigma_i(t_j))$, an acceptable solution will fulfil $\overline{y}_i(t_j) - \sigma_i(t_j) < y_i(t_j, \theta) < \overline{y}_i(t_j) + \sigma_i(t_j)$. (51)

The objective function incorporates these conditions and measures the deviation of the model from the healthy state. In the TIde approach, the objective function is as the sum of squared normalised residuals

$$f = \sum_i \left(\frac{\mu_i(t_j) - X_i(t_j)}{\sigma_i(t_j)} \right)^2 .$$

Summands of the objective function are added depending on the healthy conditions that have to be fulfilled:

- If the concentration of a species has to be minimised beyond a certain threshold, the value of $\mu_i(t_j)$ is 0, while $X_i(t_j) = y_i(t_j, \theta)$, so that the lower the concentration is, the lower is the value of f. The threshold of acceptability $\overline{y}(t_j)$ is incorporated as variance $\sigma_i(t_j)$, so that when $y_i(t_j, \theta) = \sigma_i(t_j)$, the summand is equal to 1. This is incorporated to the objective function as:
$$f = \cdots + \left(\frac{0 - y_i(t_j, \theta)}{\overline{y}_i(t_j)} \right)^2 + \ldots$$

- If the concentration of a species has to be maximised beyond a certain threshold, the value of $\mu_i(t_j)$ is 0, while $X_i(t_j) = \frac{1}{y_i(t_j, \theta)}$. Thus, the higher the concentration is, the lower is the value of f. The inverse of the threshold of acceptability $\frac{1}{\overline{y}(t_j)}$ is used as variance $\sigma_i(t_j)$, so that when $y_i(t_j, \theta) = \sigma_i(t_j)$, the summand is equal to 1. This is incorporated to the objective function as:
$$f = \cdots + \left(\frac{0 - \frac{1}{y_i(t_j, \theta)}}{\frac{1}{\overline{y}_i(t_j)}} \right)^2 + \ldots$$

- If the concentration of a species has to be within a interval $(\overline{y}_i(t_j) - \sigma_i(t_j), \overline{y}_i(t_j) + \sigma_i(t_j))$, the value of $\mu_i(t_j)$ is the central value $\overline{y}_i(t_j)$ of the interval, while $X_i(t_j) = y_i(t_j, \theta)$, so that the nearer the concentration is to $\overline{y}_i(t_j)$, the lower is the value of f. The difference $\sigma_i(t_j)$ between the limits of the interval and its centre, is used as variance, so that when $y_i(t_j, \theta) = \overline{y}_i(t_j) \pm \sigma_i(t_j)$, the summand is equal to 1. This is incorporated to the objective function as:
$$f = \cdots + \left(\frac{\overline{y}_i(t_j) - y_i(t_j, \theta)}{\sigma_i(t_j)} \right)^2 + \ldots$$

3.3.4 Determination of successful drug treatments

The objective function f takes the follows a χ^2 distribution assuming that the residuals $\left(\frac{\mu_i(t_j - X_i(t_j)}{\sigma_i(t_j)}\right)$ are independent and standard normally distributed. Therefore the similarity between the model and the healthy state could be determined by performing this statistical test with the corresponding degrees of freedom. However, as there is no reason to assume that residuals are standard normally normally distributed, another criterion for the acceptability of a drug treatment is defined.

Since the objective function has the structure of a sum of squared elements (which are always positive), it can be assured that if $f = n$, it holds that $\left(\frac{\mu_i(t_j - X_i(t_j)}{\sigma_i(t_j)}\right)^2 \leqslant n$ for every value of (i, j). Because of that, if $f \leqslant 1$ it follows that $\left(\frac{\mu_i(t_j - X_i(t_j)}{\sigma_i(t_j)}\right) \leqslant 1$ for every value of (i, j), which means that all the healthy conditions described as summands of the objective function are fulfilled. Therefore, a drug combination will be accepted as successful if $f \leqslant 1$. (51)

TIde provides several different optimisers for the simulation of drug treatments and the estimation of the drug combinations that can lead a pathological model to healthy state, including BFGS (48), simulated annealing (49), Nelder Mead (53) or a genetic algorithm (54). However, due to the high speed of simulations, brute-force optimisation is also possible. In brute-force optimisation, the value of the objective function for the whole parameter space is directly scanned for every drug through an interval, provided a maximum and minimum concentration threshold and a number of simulation steps. Brute force optimisation is very robust and can also be used for testing combinations of drug pairs for small models, while for testing for higher order combinations other more efficient algorithms such as the optimisers mentioned above have to be used. (51)

3.4 Drug target identification in healthy models

TIde provides a useful framework for the identification of drug targets in ODE models of pathological system. But in many cases, it is not possible to study an illness using this method, as models of diseases are specific to one disease. There is therefore a total dependence on the existence of pathological models. While for some cases, such as infectious diseases, models relevant to the disease might be available, this is generally the case for multifactorial diseases, as they

cannot be described by a single model. Multifactorial diseases have a great impact on human health and including disorders such as cancer, diabetes, obesity or hypertension. There might exist models which describe the healthy state of the biological system, but these models cannot be used for drug target identification using this approach, as they do not include a description of the pathological state to use as a starting point of the drug search. In this work, I adapt this framework to provide a solution for this problem of how to predict possible drug targets on models that describe the healthy state of a system, in order to study multifactorial diseases.

My approach is based on the assumption that models describing healthy systems can be used for the simulation of mutations that lead to a pathological state. This is done by systematically impairing every reaction and species in the model creating a set of perturbed models, which model potential pathological conditions. The pathological model set is evaluated for an objective function describing the characteristics of the initial healthy model which are thought to be most relevant for the functioning of the system. After this, the pathological systems are scanned for drug treatments that can restore the original value of the objective function, and thus rescue the phenotype of the model. The final output is a set of models, with different kinds of perturbations and all the drug treatments that have been able to restore the original healthy condition.

This methodological approach has been developed as a Python script (see Appendix 1), which makes use of the libTI2 libraries of the semanticSBML package.

3.4.1 *A priori* simulation of pathological models

At the start of the approach pipeline, the healthy model is manipulated in the manner described in section 3.3.2 to include the kinetics of all possible drug modifiers: inhibitors and non-essential activators for constant flux, mass action law and other kinetics and competitive inhibitors, non-competitive inhibitors, uncompetitive inhibitors and non-essential activators for Michaelis-Menten kinetics. Furthermore, an additional type of inhibitor is incorporated, that does not affect the kinetics of existing reactions, but adds new ones. These inhibitors are inactivators that specifically and reversibly bind a species of the model, producing an inactive form:

$$S + I \rightleftharpoons S_{inhibited}$$

$$k = \frac{[S][I]}{[S_{inhibited}]} = 1$$

The concentration of inactivator is modelled as a constant, so that at a every time point, if $[I] = 1$, $\frac{1}{2}$ of the targeted species are in inactive form and fulfilling that $[S] = [S_{inhibited}]$-

After all the different types of modifiers and new species have been incorporated to the model, a set of knock-out models is created. This is done by iteratively inhibiting every reaction and species in the model. Reactions are knocked-out by fixing the concentration of the corresponding (non-competitive) inhibitor to the arbitrarily high value of 99999999, which makes the reaction function at 10^{-8} of its normal velocity. Species are knocked-out in a similar way, by fixing the concentration of the corresponding inactivator to 10^8, so that only approximately 10^{-8} of the species is at free state. Additionally, when knocking-out a component of the model affects the system in an unrepairable way, softer (knock-down) conditions are employed. In these conditions, both (non-competitive) inhibitors and inactivators are fixed at a concentration of 1, so that the knocked-down reaction functions at $\frac{1}{2}$ of its normal velocity or $\frac{1}{2}$ of the knocked-down species is at an inactive state.

Since the concept of a biological species (for example, a certain protein) does not have to exactly correspond to that of the species in the model, this approach allows for inputting which species in the model are to be simultaneously knocked-out, by fixing the same values of inactivator modifiers for every one of them. Furthermore, as different number of reactions can be catalysed by the same enzyme, this system also allows for simultaneously knocking out the desired number of reactions.

3.4.2 Description of the healthy state

The healthy state of the system is described as an objective function. In this approach, the values of the time course are taken from the output of the initial model, which describes the healthy state that we want to restore. The choice of the values, i.e. which species and which time point to consider, is not trivial, but must stem from knowledge on which are the most important features of the model that are essential for the correct functioning of the system. have to be considered. This framework allows for introducing as many time points as desired and can evaluate the values of the variables at these time points as well as their integrals from $t = 0$. A balance between accurateness and simplicity must be kept at the time course generation, so that the chosen time course can sufficiently describe the healthy state that we want to restore, but without introducing unnecessary constraints.

Once the time course for optimisation has been chosen, an objective function has to be constructed. The choice of the objective function is not trivial either, since the *a priori* simulation of pathological states can affect the relevant concentrations of the system in different ways. This is an additional difficulty in comparison to the use of pathological models for drug target

predictions, as in this case only the healthy time course values are known, but not whether the concentration in the new modelled will be over the threshold or below it. Because of this, one must decide how to construct the objective function. There are three possibilities:

- Only considering variations of the measurables in one direction. In this case, the objective function will be constructed so that, for every time point, it either maximises or minimises the value of the species beyond the health-threshold. If for all time points the species already surpasses that threshold, the objective function will consider the model as healthy, even though it might be radically different from the original ODE model.

- Defining the objective function as an acceptable interval. In this case, models with great variations (regardless of the direction) in the values of the time course will be considered as pathological. However, the broadness of the acceptable interval cannot be deduced from the original ODE model, so it must be based either on literature data or arbitrarily defined (for example, as a variation of $\pm 20\%$ of the healthy value).

- Consider variations of the measurables in both directions. This framework allows for the simultaneous use of multiple objective functions. Therefore, one can define one objective function so that it maximises the values of the observables and another one for minimisation of the values for every time point.

For the assays performed in this work, I have preferred the third approach to construct the objective function, since it requires no previous assumption (or knowledge) of which interval around the original value can be considered acceptable and can be used both for models that show increased and decreased values of the observables. The idea behind the approach is the following:

A set of 2^n objective functions is constructed, where n is the number of time points in the time course. For each time point, two objective functions are defined, one that aims for the maximisation of the observables and another that aims for their minimisation. These objective functions are used for evaluating the effect of the mutation on every time point. The objective functions with a value over 1, describing the observables that have been modified by the mutation and whether they are increased or decreased, are then considered for optimisation. In this tool, I have implemented this idea for one time point and two objective functions, so that they respectively describe the minimisation and the maximisation of one observable at one time point. Therefore, only one of them will show values over 1, and will be considered for the optimisation.

3.4.3 Determination of successful drug treatments

After the objective function is defined, the program proceeds to evaluating every knock-out model in the set for the search of possible drugs that can restore the healthy state of the biological system. This is performed with a parameter estimation, where the parameter vector comprises concentrations of all modelled drugs, except for those affecting the same reaction or species that has been knocked-out. Drug scans in pathological models in this work are performed by using a brute-force 1-dimensional optimisation. A maximum and a minimum concentration for the drugs is defined and a number n of steps, so that every drug is individually assayed for n concentrations in a uniform logarithmic scale between the defined minimum and maximum. Since some reactions or molecules can be targeted by the same drug, this approach allows for inputting which reactions or species in the model share common drug affinities, so that they are targeted simultaneously by using the same concentrations of drugs.

If at least one drug treatment is found to satisfy that the objective function $f < 1$ for two consecutive time points, the simulation is considered successful and a list of drugs that satisfy this condition and their concentration at $f = 1$, is provided. If no drug treatment is successful, the simulation is re-run using the knock-down model for the same reaction/species, which shows a less marked effect on the behaviour of the system. If the simulation does not find a successful treatment for the knock-down either, the model is considered unrestorable.

4

Results

4.1 Simple pathway example

4.1.1 Description of the model

As a proof of concept for the applicability of my framework, I have created an ODE model describing a simple pathway, using the tool COPASI(19). The model is composed of 8 species and 10 reactions and consists of a linear part which bifurcates into a bi-branched pathway, joining again at the end. The linear part describes the synthesis of the species S_1 and its successive transformation into S_2 and S_3. At this point the pathway splits into two sub-pathways, the first of them comprising species S_4 and S_5 and the second, S_6 and S_7, where S_5 is a non-essential activator of $S_3 \rightarrow S_6$ and S_7 is a non-essential activator of $S_3 \rightarrow S_4$. The pathway converges at S_8, which is then degraded. All reactions in the pathway are irreversible. Seven of them follow mass-action kinetics, one is a constant flux, and the remaining two are described by user defined laws. A graphical representation of the model is shown on Figure 4.1, and their reaction kinetics can be seen on Table 4.1.

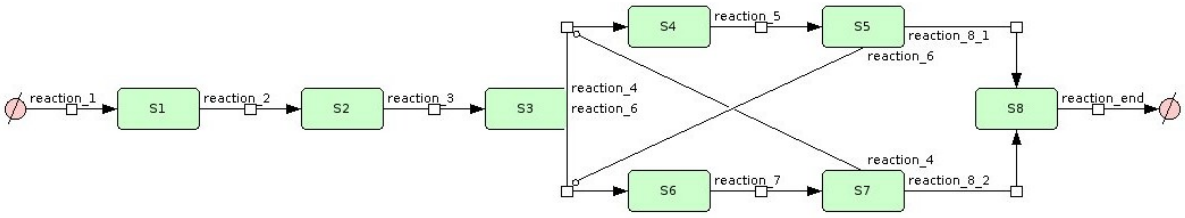

Figure 4.1: Visualisation of the example pathway - Figure generated with CellDesigner(18).

Reaction		Rate Law	Kinetic expression
1	$\rightarrow S_1$	Constant flux	$V_1 = 0.1$
2	$S_1 \rightarrow S_2$	Mass action	$V_2 = 0.1 \cdot S_1$
3	$S_2 \rightarrow S_3$	Mass action	$V_3 = 0.1 \cdot S_2$
4	$S_3 \rightarrow S_4$	Own rate law	$V_4 = 0.1 \cdot S_3 \cdot (1 + 0.1 \cdot S_7)$
5	$S_4 \rightarrow S_5$	Mass action	$V_5 = 0.1 \cdot S_4$
6	$S_3 \rightarrow S_6$	Own rate law	$V_6 = 0.1 \cdot S_3 \cdot (1 + 0.1 \cdot S_5)$
7	$S_6 \rightarrow S_7$	Mass action	$V_7 = 0.1 \cdot S_6$
8	$S_5 \rightarrow S_8$	Mass action	$V_8 = 0.1 \cdot S_5$
9	$S_7 \rightarrow S_8$	Mass action	$V_9 = 0.1 \cdot S_7$
10	$S_8 \rightarrow$	Mass action	$V_{10} = 0.1 \cdot S_8$

Table 4.1: Reactions of the example pathway - The reactions of this pathway follow mostly mass action kinetics, with the exception of reaction 1, a constant flux, and reactions 4 and 6, where the reaction rate depends both on the substrate and the products of reactions 7 and 5, respectively. For simplicity, all reactions share a common reaction constant $k = 0.1$.

Numerical simulations of the model, as visualised in Figure 4.2, show that the system reaches a steady state approximately at $t = 100$. This is possible since the only synthesis and degradation reactions in the pathway have the same rate for $S_8 = 1.0$. In this state, the concentrations of S_1, S_2 and S_8 is equal to 1, $S_4, S_5, S_6, S_7 = 0.5$ and $S_3 = 0.48$. Since the degradation of S_3 is activated by species S_5 and S_7, in steady state its concentration is kept at lower levels than that of its substrate and products.

4.1.2 Results

Applying my framework to this example pathway results in a set of knock-out and knock-down versions of the model, where either one species or one reaction is inhibited. As described in Section 3.4.3, knock-down models are only created in case the corresponding knock-out is not restorable by any possible treatment. The change of S_8 concentration at an arbitrarily large ($t = 500$) time point was chosen as the reference for assessing the influence of perturbations on the steady-state values of the system. The information on the mutant models is shown on Table 4.2.

A total of 25 mutant models were generated of which 17 showed a change of $S_{8(t=500)}$. These models are classified as pathological, as they could theoretically describe an illness stemming from alterations in the pathway. A drug scan was performed for the 15 pathological models,

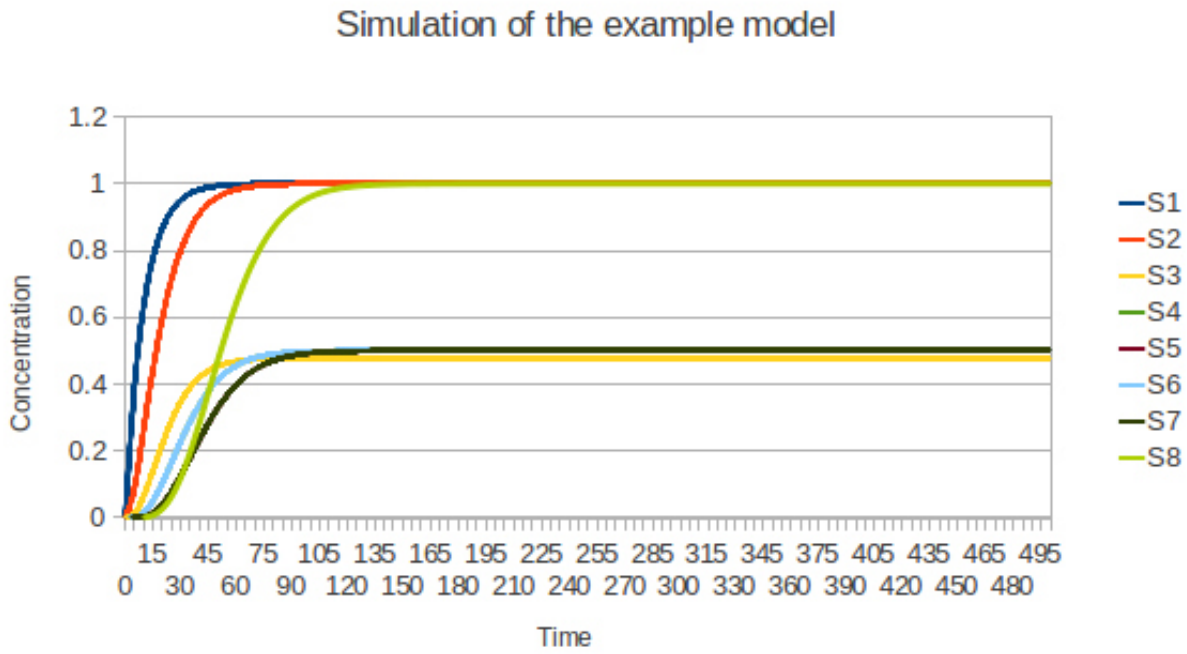

Figure 4.2: Simulation of the behaviour of the system - Simulation performed with COPASI (LSODA(45) solver). Figure generated with LibreOffice Calc.

Model name	**Change of S_8 ($t = 500$)**
Knock-out reaction 1 ($KOre_1$)	-100.00%
Knock-down reaction 1 ($KDre_1$)	-50.00%
Knock-out reaction 2 ($KOre_2$)	-100.00%
Knock-down reaction 2 ($KDre_2$)	0.0%
Knock-out reaction 3 ($KOre_3$)	-100.00%
Knock-down reaction 3 ($KDre_3$)	0.0%
Knock-out reaction 4 ($KOre_4$)	0.0%
Knock-out reaction 5 ($KOre_5$)	-51.19%
Knock-out reaction 6 ($KOre_6$)	0.0%
Knock-out reaction 7 ($KOre_7$)	-51.19%
Knock-out reaction 8 ($KOre_8$)	-28.77%
Knock-out reaction 9 ($KOre_9$)	-28.77%
Knock-out reaction 10 ($KOre_10$)	4452.38%
Knock-out S_1 (KOS_1)	-100.00%
Knock-down S_1 (KDS_1)	0.0%
Knock-out S_2 (KOS_2)	-100.00%
Knock-down S_2 (KDS_2)	0.0%
Knock-out S_3 (KOS_3)	-100.00%
Knock-down S_3 (KDS_3)	0.0%
Knock-out S_4 (KOS_4)	-51.19%
Knock-out S_5 (KOS_5)	-51.19%
Knock-out S_6 (KOS_6)	-51.19%
Knock-out S_7 (KOS_7)	-51.19%
Knock-out S_8 (KOS_8)	-100.00%
Knock-down S_8 (KDS_8)	0.0%

Table 4.2: Simulated mutants for the example model - Bold are those models for which at least one treatment could restore the original value of S_8 ($t = 500$)

using two objective functions:

$$f = \left(\frac{S_{8(t=500)}}{\overline{S}_{8(t=500)}} \right)^2$$

$$f = \left(\frac{\overline{S}_{8(t=500)}}{S_{8(t=500)}} \right)^2$$

where the first objective function evaluates whether the concentration of S_8 is over 1.0, while the second evaluates whether $S_8 < 1.0$. $\overline{S}$ is the value of $S_{8(t=500)}$ for the original model $\overline{S}_8(t = 500) = 1.0$. Inhibitors and non-essential activators were simulated for each reaction and inactivators for every species. A brute force optimisation (min: 0.01, max: 10000 steps: 50) for drug concentrations was performed. For 10 of the pathological models at least one drug was found to restore the original phenotype for a concentration between 0.1 and 10000.

The results of my analysis are shown on Tables 4.1 and 4.3. The first noteworthy characteristic is that all mutant models show a downregulation of S_8 output, except for $KOre_{10}$. This is expected, as all species and reactions in the model participate in the synthesis of S_8 with the exception of $KOre_{10}$, which describes the degradation of S_8. As the effect of mutations is opposite in these two groups, one would expect that drugs that can restore the healthy phenotype in one group cannot restore it for the other, and vice versa. This is the case in this example, and as a consequence of that, I can classify drugs for their capability of restoring healthy behaviour in either of the two mutant groups (see Figure 4.3)

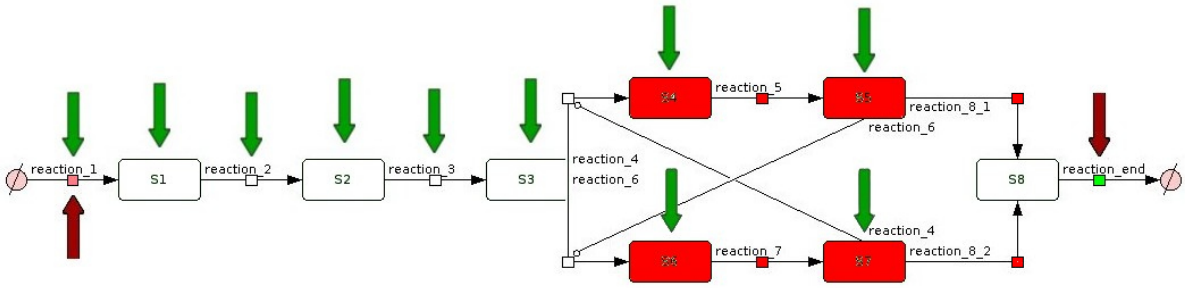

Figure 4.3: Summary of the results - Shown in red are the species or reactions whose mutation cause a downregulation of S_8 output, in green the species or reactions whose mutation cause an upregulation. Lighter shades indicate that the model is a knock-down while darker shades indicate knock-out models. The arrows symbolise the target point of successful drug treatments, with red arrows restoring red mutants and green arrows, green mutants. Arrows pointing downwards indicate inhibitors (for reactions) or inactivators (for species), while arrows pointing upwards indicate non-essential activators (for reactions).

Model name	Successful drugs (effective concentration)
$KDre_1$	reaction_{10} inhibitor (1.03)
$KOre_5$	reaction_1 activator (1.14)
	reaction_{10} inhibitor (1.09)
$KOre_7$	reaction_1 activator (1.14)
	reaction_{10} inhibitor (1.09)
$KOre_8$	reaction_1 activator (0.35)
	reaction_{10} inhibitor (0.40)
$KOre_9$	reaction_1 activator (2.13)
	reaction_{10} inhibitor (1.09)
$KOre_{10}$	S_8 inactivator (49.75)
	reaction_1 inhibitor (50.20)
	S_4 inactivator (487.98)
	S_6 inactivator (487.98)
	S_5 inactivator (496.54)
	S_7 inactivator (496.54)
	S_1 inactivator (1216.25)
	S_2 inactivator(1216.25)
	reaction_2 inhibitor (1210.86)
	reaction_3 inhibitor (1210.86)
	S_3 inactivator (2387.29)
KOS_4	reaction_1 activator (1.14)
	reaction_{10} inhibitor (1.09)
KOS_5	reaction_1 activator (1.14)
	reaction_{10} inhibitor (1.09)
KOS_6	reaction_1 activator (1.14)
	reaction_{10} inhibitor (1.09)
KOS_7	reaction_1 activator (1.14)
	reaction_{10} inhibitor (1.09)

Table 4.3: Successful treatments for each pathological model

Two reactions determine the overall flux through the system. Reactions 1 and 10, the synthesis of the first component of the pathway and the degradation of the last, respectively, are vital for modulating the activity of the pathway, as proved by the fact that out of eight models with downregulated S_8 seven can be restored by a non essential activator of reaction 1 or an inhibitor of reaction 10. If this model represented a real pathway, the results would suggest concentrating on these two reactions for the search of possible drug targets, since they can restore almost all possible alterations in the network. In this search, the focus should be especially put on reaction 10, as inhibitors are easier to construct than activator drugs.

For the only model ($KOre_10$) with upregulated S_8, multiple drugs have been found to be able to restore the initial state. But this can be misleading, because this model does not reach a steady state, since the degradation flux is impaired, so that it is not possible that for $t = 500$ $V_1 = V_{10}$. Drugs can diminish the concentration of S_8, but do not prevent its accumulation, so the necessary treatments depend on the chosen time point. This phenomenon, the disappearance of steady state due to mutations, should be taken into account when elaborating the objective function for a real analysis to ensure that the basic behaviour of the model is maintained and not only the value at one time point. Additionally, for high concentrations the model was found to be restorable by drugs that theoretically should not have any impact downregulating S_8, such as reaction 6 or reaction 9 non-essential activators. As these results could not be reproduced on COPASI and only appeared for certain tolerance parameters of the solver, I determine that these results are artifacts of the numerical optimisation and did not take them into account.

An interesting observation is that mutations on reactions and on species have different effects. For example, $KOre_4$ does not show any change on S_8 output, while $K0S_4$ clearly has S_8 downregulated. $KDre_1$ shows a reduction in 50% of output, while KDS_1 does not. These differences stem from the fact that measures are taken at a $t = 500$, when the system has already reached its steady state. For the case of $KOre_4$ and KOS_4, knocking out reaction 4 produces a slowing-down of the production of S_8, but the flux of concentrations is finally deviated through the other branch. This is not possible for KOS_4 because a only a part of the S_3 is converted into S_4, while the rest enters into a dead-end, accumulating as inactive S_4, so that the system cannot reach the steady state. Knocking-down S_1 causes a temporal descent of flux through the pathway, but as reaction 2 depends on the concentration of S_1, it is accumulated, thus overcoming the inactivation of half of the species and reaching the steady state. However, knocking down reaction 1 directly affects the steady state, as it is the only reaction that increments the total concentration of species.

The fact that mutations on reactions and on species do not have the same impact highlights the need of treating metabolic pathways and signalling pathways differently. Metabolic pathways, on the one hand, consist of a series of metabolites, connected by reactions which are catalysed by enzymes (proteins). These enzymes are encoded in the genome, so that mutations in this context affect primarily the reactions. Signalling pathways, on the other hand, consist of interactions between proteins, so that mutations would affect in principle the species, and in some cases the reactions, as a mutation can affect the interaction between species (i. e. changing the binding site of a protein). Therefore, biological perturbations in metabolic pathways will be more adequately modelled by reaction knock-outs and knock-downs, while in signalling pathways, species (and some reaction) mutants will model possible pathological conditions.

4.2 Implementation for a biologically relevant example: TGF-β signalling pathway

As a biological application for my tool, I have investigated TGF-β signalling with the objective of finding possible pathologies produced by perturbations in this biological network and treatments that can restore them. I have chosen this pathway due to its complex relationship with cancer and other diseases in human beings (see Section 1.3.2) and the availability of computational models. In my considerations, the models by Zi *et al.* 2011(34) and Wegner *et al.* 2012(25) have been used as the reference framework for the study of this pathway, because of their novelty and their level of detail. Both models are available on BioModels (14).

4.2.1 Description of the Zi *et al* 2011 TGFβ model

The TGF-β model of Zi *et al.* 2011 is a computational model of TGF-β pathway on human keratinocytes (HaCaT), created with the objective of describing the different responses of cells to to continuous and pulsating TGF-β stimulation. The model consists of 21 species, 3 compartments and 29 reactions, most of which follow mass action kinetics.(34)

The activation of the signalling route by TGFβ results in the accumulation of Smad proteins in the nucleus, of which the phosphorylated dimers Smad2-Smad4 and Smad3-Smad4 (not included) act as secondary messengers of the signalling pathway binding to the target genes and regulating their transcription, both positively and negatively. Because of this, nuclear Smad2-Smad4 ($PSmad2\_Smad4\_n$) is used as the output of the pathway. As it has been suggested that long-term signalling effects might be of critical importance for cell fate determination(34),

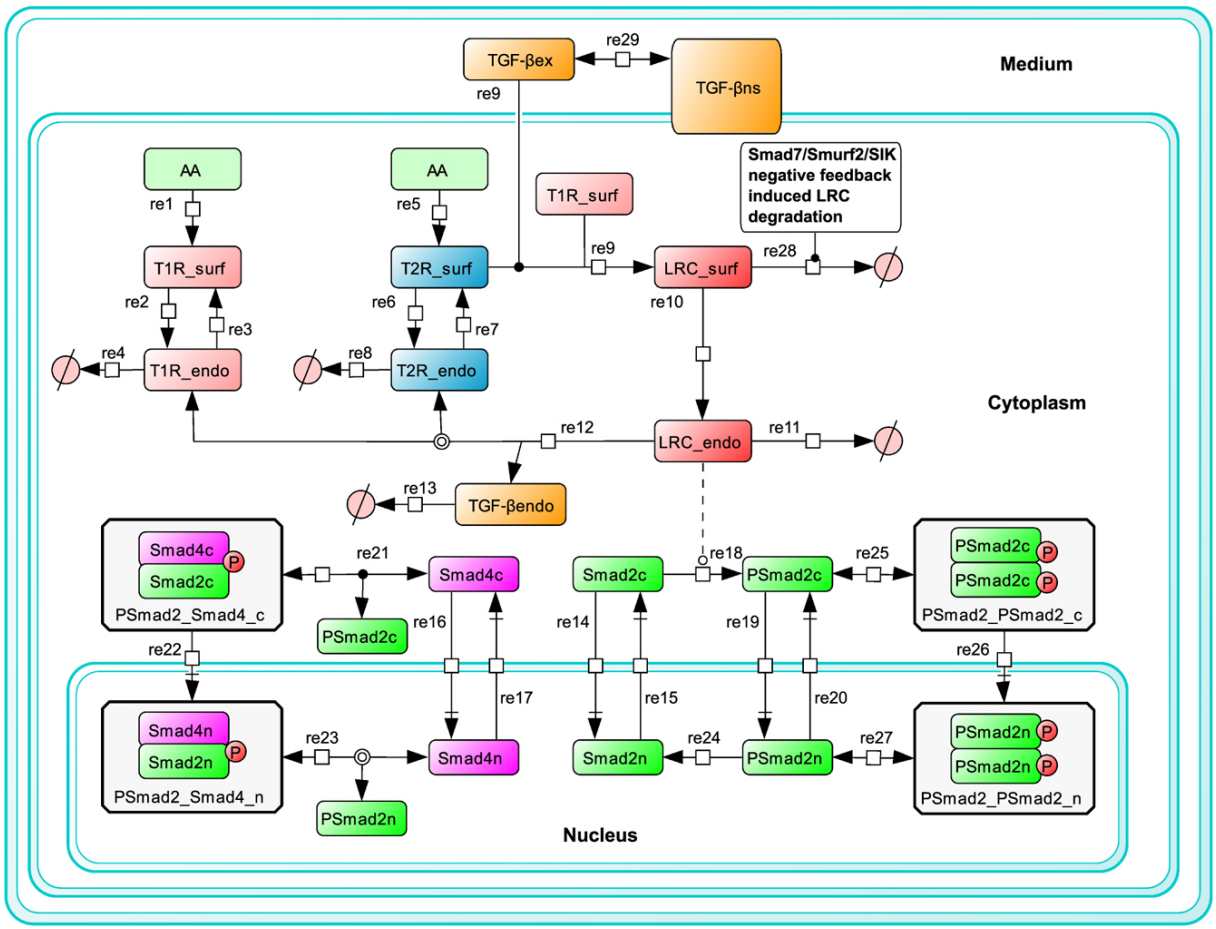

Figure 4.4: Visualisation of Zi *et al* 2011 TGFβ model - This model describes the binding of TGF-β ($TGF-$) to its two receptor molecules $T1R\_surf$ and $T2R\_surf$, forming the large receptor complex (LRC_surf). This complex is internalised and activated, catalysing the phosphorylation of $Smad2c$. Phosphorylated Smad2 ($PSmad2c$) forms heteromeric complexes with Smad4 ($PSmad2\_Smad4\_c$), as well as homomeric complexes ($PSmad2\_PSmad2\_c$). These complexes are imported and accumulate in the nucleus, as the species $PSmad2\_Smad4\_n$ and $PSmad2\_Smad4\_c$, which trigger a negative feedback-loop ($reaction28$), which degrades LRC_surf. Taken from Zi *et al* 2011(34)

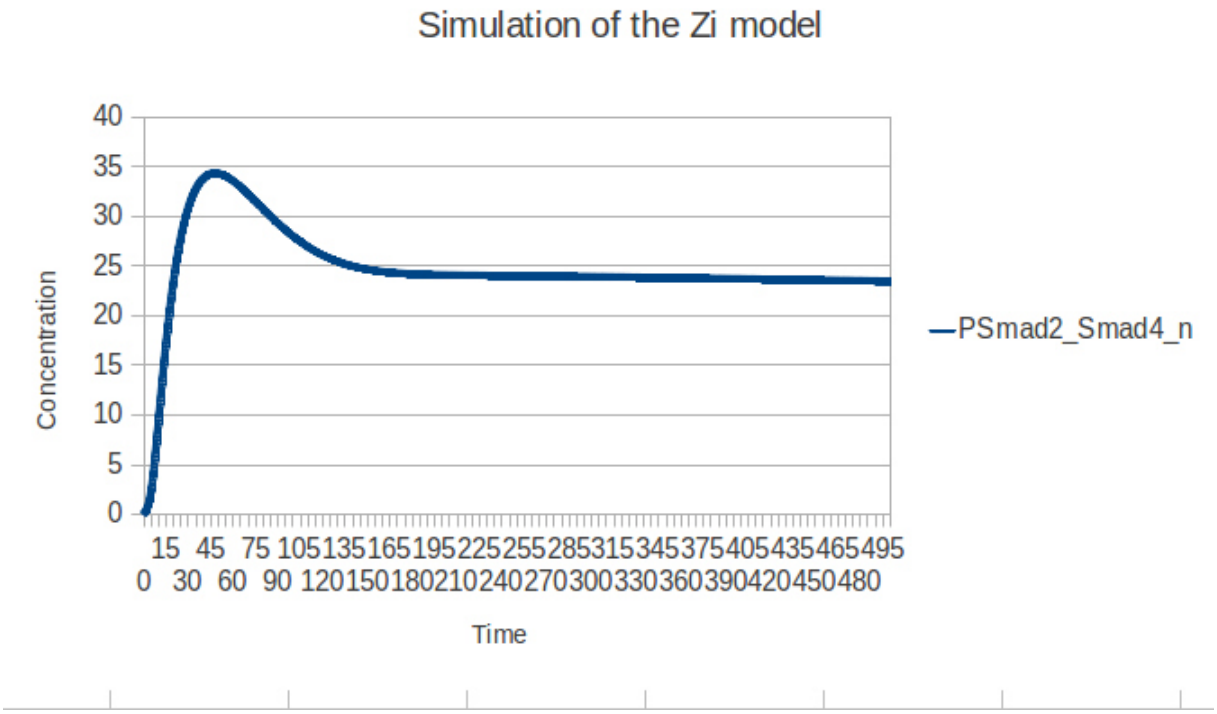

Figure 4.5: Time course of $PSmad2\_Smad4\_n$ **(**nM**) vs. time (**min**) for** $TGF - \beta = 0.05nM$ - Simulation performed with COPASI (LSODA(45) solver). Figure generated with LibreOffice Calc.

I have used a long time for the time course. This model reaches a quasi-steady state of $PSmad2\_Smad4\_n \approx 24.0$ nM at $t > 150$ min for an initial concentration of TGF-β of 0.05 nM (see Figure 4.5). Therefore, I have constructed the following objective functions for the description of the healthy state, taking as reference time point the arbitrarily high $t = 500$ min:

$$\begin{aligned} f &= \left(\frac{PSmad2\_Smad4\_n_{(t=500)}}{\overline{PSmad2\_Smad4\_n}_{(t=500)}} \right)^2 = \left(\frac{PSmad2\_Smad4\_n_{(t=500)}}{2.33} \right)^2 \\ f &= \left(\frac{\overline{PSmad2\_Smad4\_n}_{(t=500)}}{PSmad2\_Smad4\_n_{(t=500)}} \right)^2 = \left(\frac{2.33}{PSmad2\_Smad4\_n_{(t=500)}} \right)^2 \end{aligned} \tag{4.1}$$

where $PSmad2\_Smad4\_n_{(t=500)}$ is the output for the healthy model.

4.2.2 Results

Applying my framework to the Zi *et al* 2011 TGFβ model results in a set of knock-out and knock-down versions of the model, where either one species or one reaction is inhibited. The information on the mutant models is shown on Table 4.2. A total of 43 mutant models were generated (see Table 4.4).

A drug scan was performed for them, using the objective functions as shown in Eq. 4.1 and a correspondences file describing the which species are knocked-out together, because they are forms of the same protein (see Appendix 2). Using a brute force optimisation (min: 0.01, max: 10000, steps: 50) at least one drug was found to restore the original phenotype for 33 of the models. For simplicity, the obtained results were plotted as a matrix and are shown on Figure 4.6.

Both, models and drugs, can be divided roughly into two groups. Each group of drugs can restore one group of mutant models but not the other. This behaviour strongly resembles that of the example pathway, where drugs could be classified between those that restore downregulated models and those that restore upregulated models. However, in this case there are models that are unexpectedly restored by drugs that do not have an action on similar models. For example, $KDT2R$, a downregulated model with a $\Delta PSmad2\_Smad4\_n_{(t=500)} = -0.10\%$, can be restored by $PSmad2n\_inactivator$, which also restores other upregulated models including $KOre28$, with a $\Delta PSmad2\_Smad4\_n_{(t=500)} = 172.36\%$. This feature could stem from the complexity of the biochemical network so that the same drug can have different effects in different models due to interactions between the mutated gene and the drug target. But as every model, regardless of

Model name	Change of $PSmad2\_Smad4\_n$ ($t = 500$)
Knock-out reaction 1 ($KOre_1$)	−76.34%
Knock-out reaction 2 ($KOre_2$)	6.15%
Knock-out reaction 3 ($KOre_3$)	−30.33%
Knock-out reaction 4 ($KOre_4$)	4.06%
Knock-out reaction 5 ($KOre_5$)	−99.07%
Knock-out reaction 6 ($KOre_6$)	6.13%
Knock-out reaction 7 ($KOre_7$)	−13.82%
Knock-out reaction 8 ($KOre_8$)	7.20%
Knock-out reaction 9 ($KOre_9$)	−100.00%
Knock-down reaction 9 ($KDre_9$)	−3.68%
Knock-out reaction 10 ($KOre_{10}$)	−100.00%
Knock-down reaction 10 ($KDre_{10}$)	−24.03%
Knock-out reaction 11 ($KOre_{11}$)	3.32%
Knock-out reaction 12 ($KOre_{12}$)	163.46%
Knock-out reaction 13 ($KOre_{13}$)	0.00%
Knock-out reaction 14 ($KOre_{14}$)	8.54%
Knock-out reaction 15 ($KOre_{15}$)	−100.00%
Knock-down reaction 15 ($KDre_{15}$)	−7.14%
Knock-out reaction 16 ($KOre_{16}$)	−55.51%
Knock-out reaction 17 ($KOre_{17}$)	−58.79%
Knock-out reaction 18 ($KOre_{18}$)	−100.00%
Knock-down reaction 18 ($KDre_{18}$)	−27.62%
Knock-out reaction 19 ($KOre_{19}$)	0.32%
Knock-out reaction 20 ($KOre_{20}$)	−5.74%
Knock-out reaction 21 ($KOre_{21}$)	−13.89%
Knock-out reaction 22 ($KOre_{22}$)	−34.06%
Knock-out reaction 23 ($KOre_{23}$)	474.52%
Knock-out reaction 24 ($KOre_{24}$)	234.07%
Knock-out reaction 25 ($KOre_{25}$)	0.17%
Knock-out reaction 26 ($KOre_{26}$)	0.15%
Knock-out reaction 27 ($KOre_{27}$)	−23.23%
Knock-out reaction 28 ($KOre_{28}$)	172.36%
Knock-out reaction 29 ($KOre_{29}$)	0.11%
Knock-out $Smad2$ ($KOSmad2$)	−100.00%
Knock-down $Smad2$ ($KDSmad2$)	−29.50%
Knock-out $Smad4$ ($KOSmad4$)	−100.00%
Knock-down $Smad4$ ($KDSmad4$)	−31.51%
Knock-out $TGF-\beta$ ($KOTGF-\beta$)	−100.00%
Knock-down $TGF-\beta$ ($KDTGF-\beta$)	−3.68%
Knock-out $T1R$ ($KOT1R$)	−100.00%
Knock-down $T1R$ ($KDT1R$)	−0.47%
Knock-out $T2R$ ($KOT2R$)	−100.00%
Knock-down $T2R$ ($KDT2R$)	−0.10%

Table 4.4: Simulated mutants for the Zi *et al* 2011 TGFβ model - Bold are those models for which at least one treatment could restore the original value of $PSmad2\_Smad4\_n$ ($t = 500$)

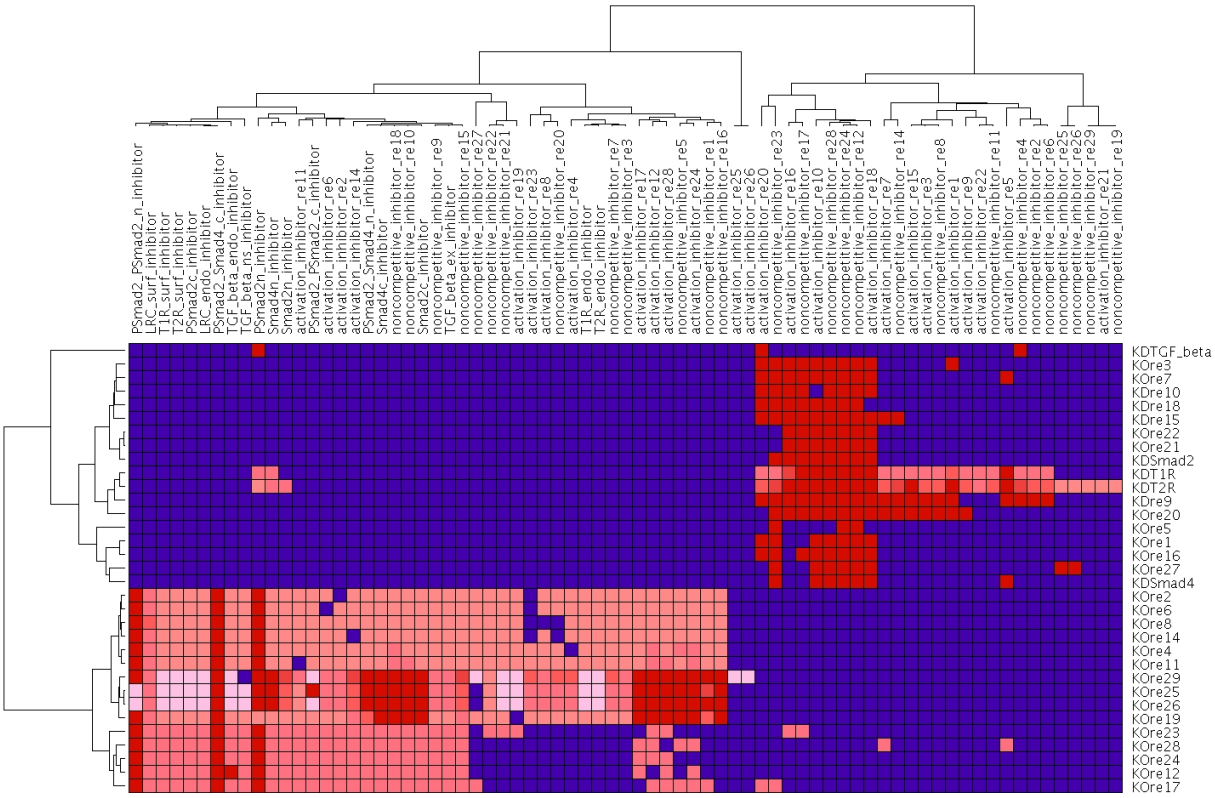

Figure 4.6: Successful treatments for mutant models of Zi *et al* 2011 TGFβ model - Each cell describes the combination of a drug (columns) with a mutant model (rows). Shown in blue are unsuccessful treatments and treatments that could restore healthy conditions on different shades of red. Darker shades of red indicate drugs effective at lower concentrations, lighter shades indicate drugs effective at higher concentrations. Hierarchical clustering of the drugs and models was performed using pairwise-complete linkage method and Pearson correlation as a distance measure. Analysis and visualisation have been performed in generated with GenePattern(55)

their output change, is included in the analysis, drugs with very weak and non-linear effects may appear to be effective for mutants of the other group if they are very near to normal conditions. To exclude this effect of the use of all models in the analysis, as well as for simplifying results, I evaluated the behaviour of the models to define a *threshold of pathogeny*, a threshold to define which models are regarded as pathological or similar enough to the healthy state.

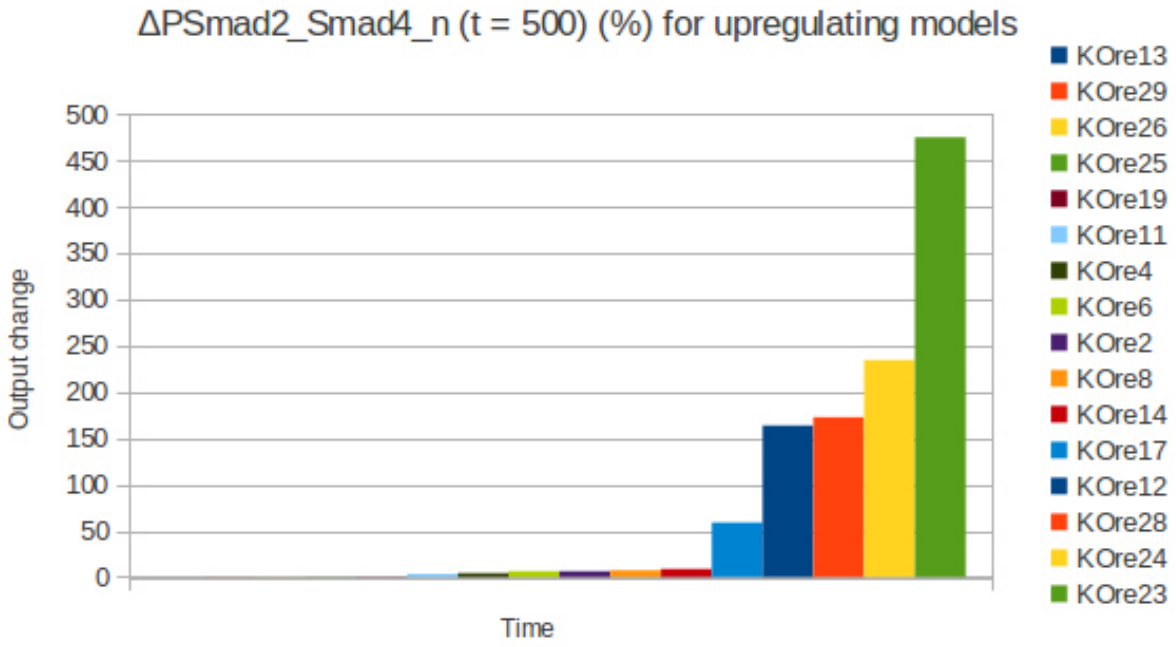

Figure 4.7: $\Delta PSmad2\_Smad4\_n_{(t=500)}$ **(%) for upregulated models** - This Figure shows the change in output for the set of upregulated mutant models, ordered from the lowest output to the highest. The figure generated has been generated with LibreOffice Calc.

As it can be observed on Figure 4.7, most of the models with upregulated TGF-β signalling show a $\Delta PSmad2\_Smad4\_n_{(t=500)}$ lower than 50%. Only four models have a output change higher than 100%, and a marked difference between the fourth and the fifth models with highest output change, and between the fifth and the sixth, is observed. As it can be seen on Figure 4.8, for the models with downregulated TGF-β signalling there are three models with a output change lower than -40% and a set of seven models with very similar output change, between 30% and 20%. Therefore, there are two possible options for defining the threshold of disease: the stricter threshold defining as diseased those models with output change higher than 100% or lower than -40%, and a less demanding threshold, defined as $\Delta PSmad2\_Smad4\_n_{(t=500)} >$

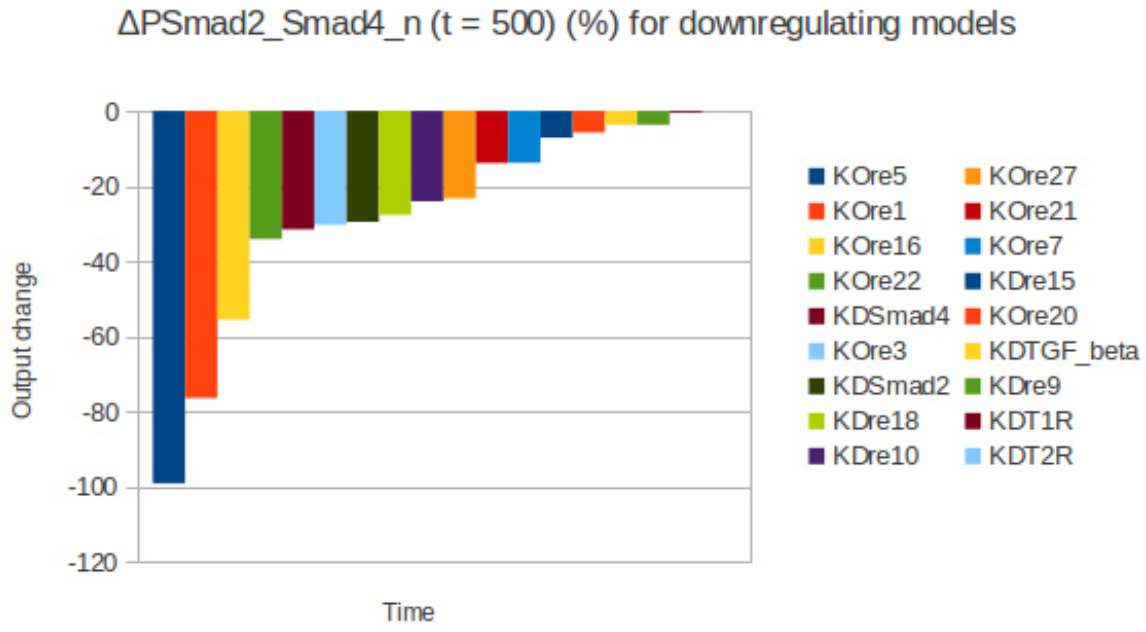

Figure 4.8: $\Delta PSmad2\_Smad4\_n_{(t=500)}$ **(%) for downregulated models** - This Figure shows the change in output for the set of downregulated mutant models, ordered from the lowest output to the highest. The figure generated has been generated with LibreOffice Calc.

$50\% \vee \Delta PSmad2\_Smad4\_n_{(t=500)} < -10\%$. To maximise the number of analysed cases, I decided to employ the second threshold.

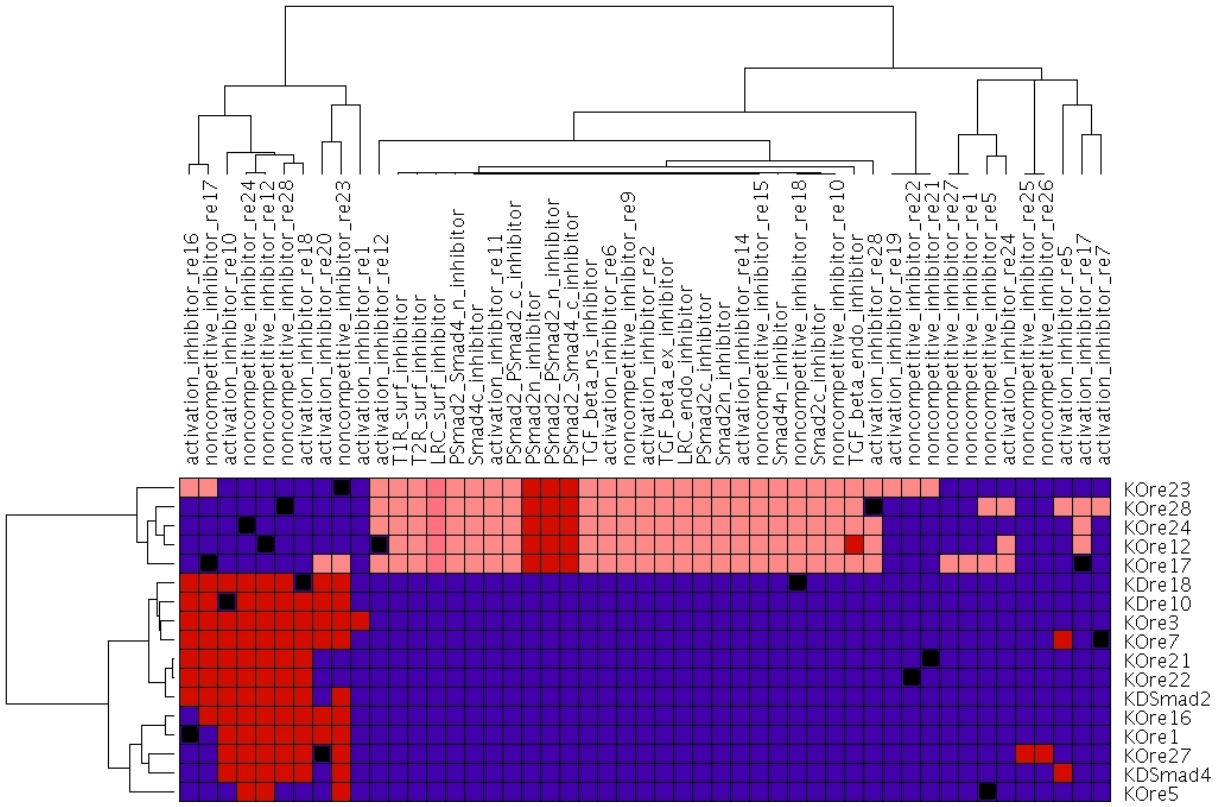

Figure 4.9: Successful treatments for pathological models of Zi *et al* 2011 TGFβ model TGFβ model - Similar analysis as in Figure 4.6. Cells in black indicate drugs which act on elements that are knocked out or knocked down in the model. Analysis and visualisation have been performed in generated with GenePattern(55) and manually modified to include those combinations that have not been assayed.

The results of the filtered analysis are shown on Figure 4.9. It can be observed that there are still some drugs with the effect of restoring both upregulated (upper part of the matrix) and down-regulated models (lower part) for some cases, such as *activation_inhibitor_re*16, *noncompetitive_inhibitor_re*17, *activation_inhibitor_re*20, *noncompetitive_inhibitor_re*23 and *activation_inhibitor_re*5. As these results could be reproduced using COPASI, I determined that they are caused by changes in the interaction between the drug target and the mutation. For example, *activation_inhibitor_re*16 and *noncompetitive_inhibitor_re*17, which in normal conditions induce the accumulation of *Smad*4 in the nucleus and therefore an augment of the output, have the opposite effect in the case of $KO23$. This can be explained by the fact that

nuclear *Smad*4 cannot dimerise, which downregulates the output. But, apart from these exceptions, drugs can be generally classified as being able to restore one group of models.

Upregulated models can generally be restored using inactivators for species in the signalling route, such as the receptors and Smad compounds. Understandably, those that directly target nuclear Smad compounds have a stronger effect on the output. They can also be restored by inhibitors of the reactions necessary for output generation, such as reactions 9, 10 and 18, and activators of the degradation of key species (i. e. reactions 11 and 18) and of the reactions that stop output generation (i. e. reactions 2, 6 and 12). Strangely, the inactivator of TGF_β_endo seems to have an effect on this model group, although this species is not the substrate of any reaction and its accumulation does not affect in any way the system. This result could not be reproduced using COPASI. Therefore, I considered this result to be an artifact of the optimisation.

For downregulated models, the list of possible drugs is shorter. $KOre5$, the model with the lowest output can only be restored by three possible drugs. This list of drugs includes activators of reactions key for signalling, i.e. reactions 10, 18 and 23, and inhibitors of those that describe degradation of important species, i.e. reactions 24, 12 and 28. The results are summarised on Figure 4.10, which shows the mutant models that have a greater change of $PSmad2\_Smad4\_n$ and the drugs that can restore them.

4.2.3 Description of the Wegner *et al* 2012 TGFβ model

The TGFβ model of Wegner is the second biologically relevant model that has been analysed using my framework. It describes TGF-β signalling on mouse hepatome cells (AML12) and has been created with the objective of describing the different feedback loops in the model and analysing their impact in the overall signalling in detail. Moreover, the model determines that for a realistic set of parameter values, TGF-β signalling leads to an oscillating response, in contrast to the Zi model. The model consists of 53 species, 2 compartments and 91 reactions, and the majority of the reactions follow mass action kinetics whereas phosphorylation and dephosphorylation of *RSmad* are described by Michaelis-Menten type kinetics.(25)

In these model, activation of the signalling route by TGFβ results in the accumulation of phosphorylated Smad2-Smad4 and Smad3-Smad4 dimers ($pRSmad\_Smad4\_n$) in the nucleus, triggering gene expression, which in this model is explicitly described as *geneProduct*. As this is the final output of the pathway, *geneProduct* is used as the species of reference for quantifying TGF-β signalling. In order to construct the objective function, I select a late time point as in

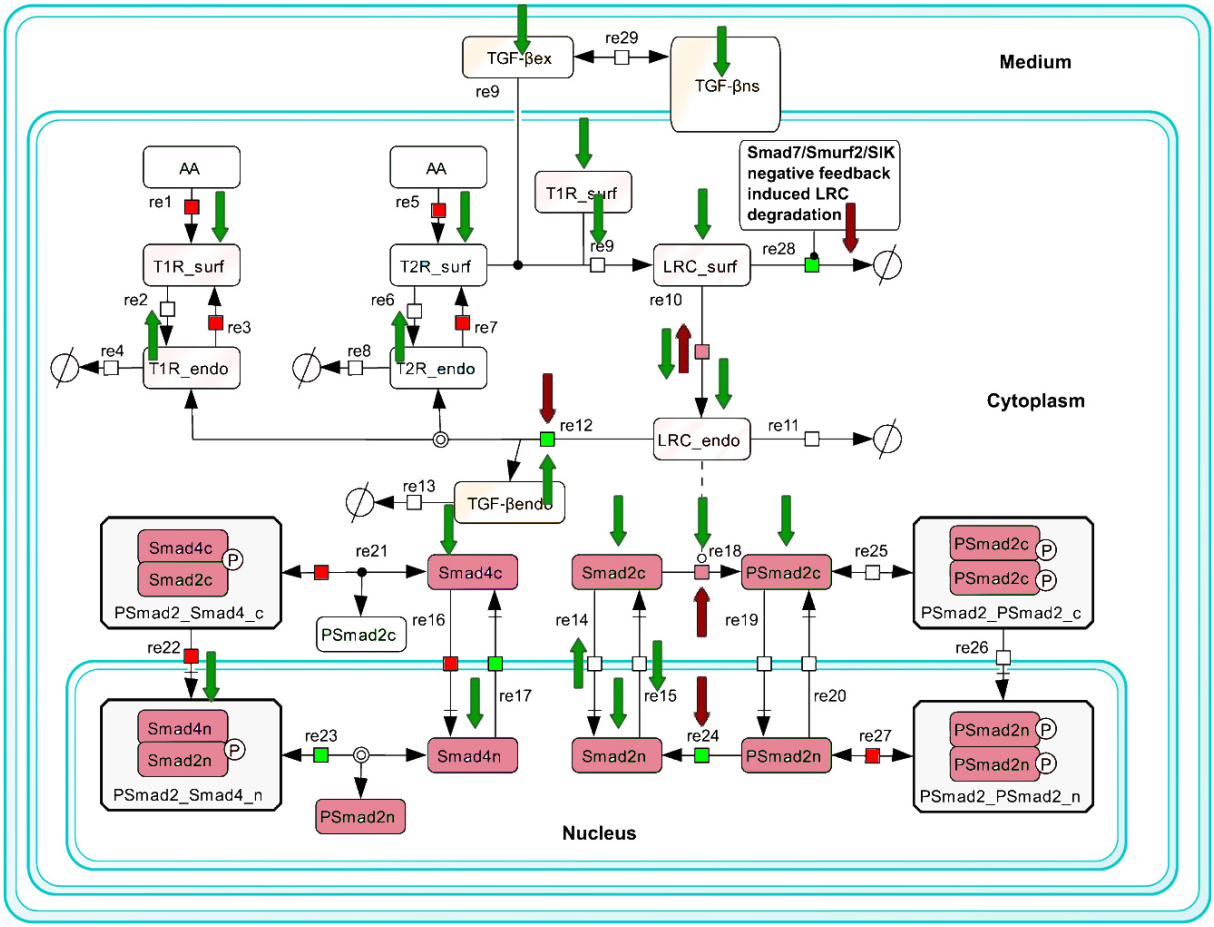

Figure 4.10: Summary of the results for Zi *et al* 2011 TGFβ model - Same analysis as in Figure 4.3

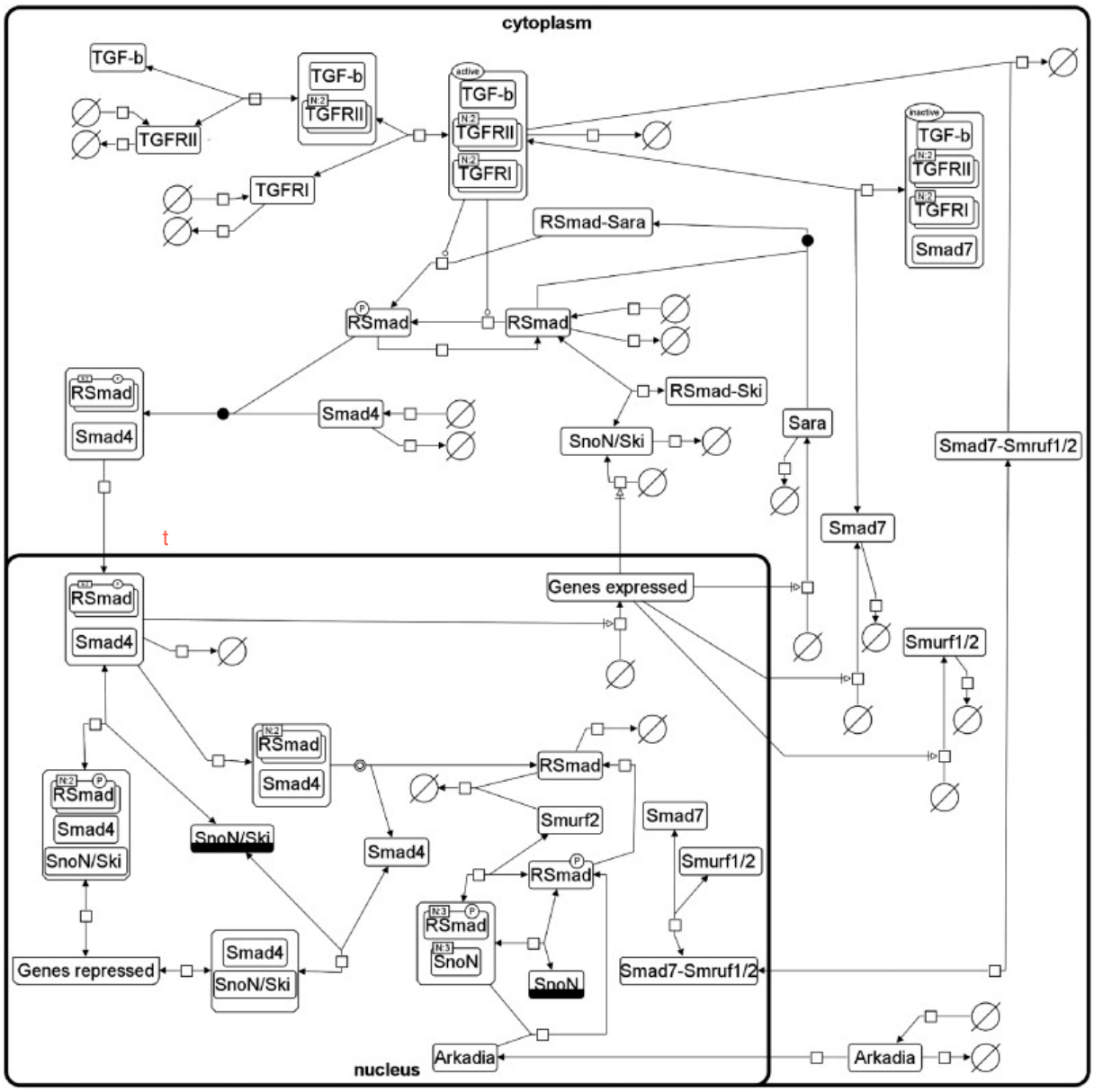

Figure 4.11: Summary of Wegner *et al* 2012 TGFβ model - This model describes the binding of TGF-β to its two receptor molecules $TGF\beta RI$ and $TGF\beta RI$. This complex is internalised and activated, catalysing the phosphorylation of Smad2 and Smad3 ($RSmad_c$), which can be facilitated by the Smad anchor for receptor activation ($SARA$). Phosphorylated RSmad ($pRSmad_c$) forms heteromeric complexes with Smad4 ($pRSmad\_Smad4\_c$). These complexes are imported and accumulate in the nucleus as the species $pRSmad\_Smad4\_n$ where they trigger gene expression. This also activates multiple feedback pathways: $SARA$, a positive feedback mediator and $Smad7$, $Smurf1/2$, Ski and $SnoN$ mediate negative feedback reactions. $Smurf2$ has an additional effect as a promoter of nuclear $SnoN$ degradation when $SnoN$ is bound to $pSmad2$ (not shown in the Figure). Additionally, the model includes $Arkadia$, a protein that degrades $Smad7$ and $SnoN$ bound to $RSmad$. This Figure has been taken from Wegner *et al* 2012(25)

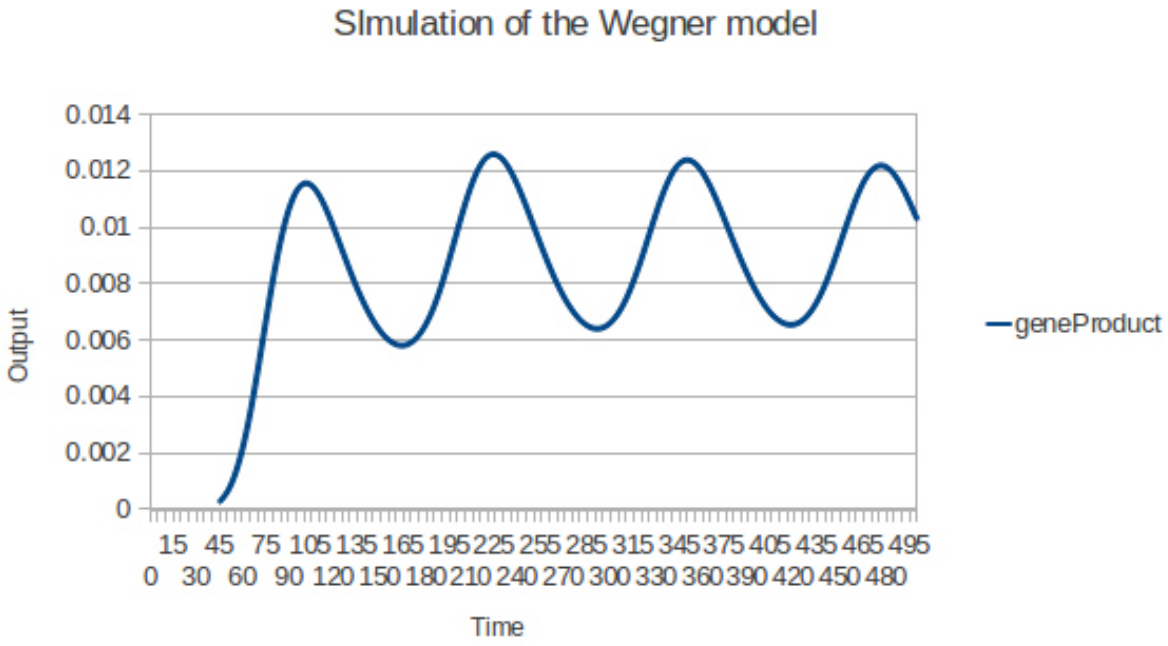

Figure 4.12: Time course of $geneProduct$ **(**μM**) vs. time (**min**) for** $TGF-\beta = 0.01\mu M$ - Simulation performed with COPASI (LSODA(45) solver). Figure has been generated with LibreOffice Calc.

the analysis of Zi *et al.* 2011 model. But since this model shows a marked oscillation for $t > 100$ min of approximately 0.009 ± 0.003 (see Figure 4.12), choosing a unique time point could lead to a shift on the oscillation being considered as an upregulation of the model. Therefore, I have constructed the objective function as a function of the integral of $geneProduct$, taking as reference time point the arbitrarily high $t = 500$ min:

$$\begin{aligned} f &= \left(\frac{geneProduct\_int_{(t=500)}}{\overline{geneProduct\_int}_{(t=500)}} \right)^2 = \left(\frac{geneProduct\_int_{(t=500)}}{3.99} \right)^2 \\ f &= \left(\frac{\overline{geneProduct\_int}_{(t=500)}}{geneProduct\_int_{(t=500)}} \right)^2 = \left(\frac{3.99}{geneProduct\_int_{(t=500)}} \right)^2 \end{aligned} \tag{4.2}$$

where

$$geneProduct\_int_{(t=500)} = \int_0^{t=500} geneProduct(t)dt$$

and $\overline{geneProduct\_int}_{(t=500)}$ is the value for the healthy model.

4.2.4 Results

A total of 120 mutant models were generated and a drug scan was performed, using the objective functions 4.2 and a correspondences file (see Appendix 2). After a linear brute force optimisation (min: 0.01, max: 10000, steps: 50) 105 of the models were found to be restorable. The information on the mutant models is summarised on Figures 4.13 and 4.14 and the results of the drug scan are shown on Figure 4.15

The results obtained from the Wegner *et al* 2012 TGFβ model are of great complexity due to the size of the model. Therefore, to simplify the output and ensure that only the drugs with strong effects are taken into account, I evaluated the behaviour of the models to again define a *threshold of pathogeny.*

As it can be observed on Figure 4.13, most of the models with upregulated TGF-β signalling show a $\Delta geneProduct\_int_{(t=500)}$ lower than 35%. Only five models have a output change higher than 100%, and there is great difference between the fifth and the sixth models and between the 17th and the 18th models with highest output change. For the case of downregulated models (Figure 4.14), there is a big shift in downregulation between models 5 and 6 and between models 9 and 10. Therefore, there are again two possible options for defining the threshold of pathogeny: the stricter threshold $\Delta geneProduct\_int_{(t=500)} > 100\% \vee \Delta geneProduct\_int_{(t=500)} < -90\%$ and a less demanding threshold, $\Delta geneProduct\_int_{(t=500)} > 35\% \vee \Delta geneProduct\_int_{(t=500)} <$

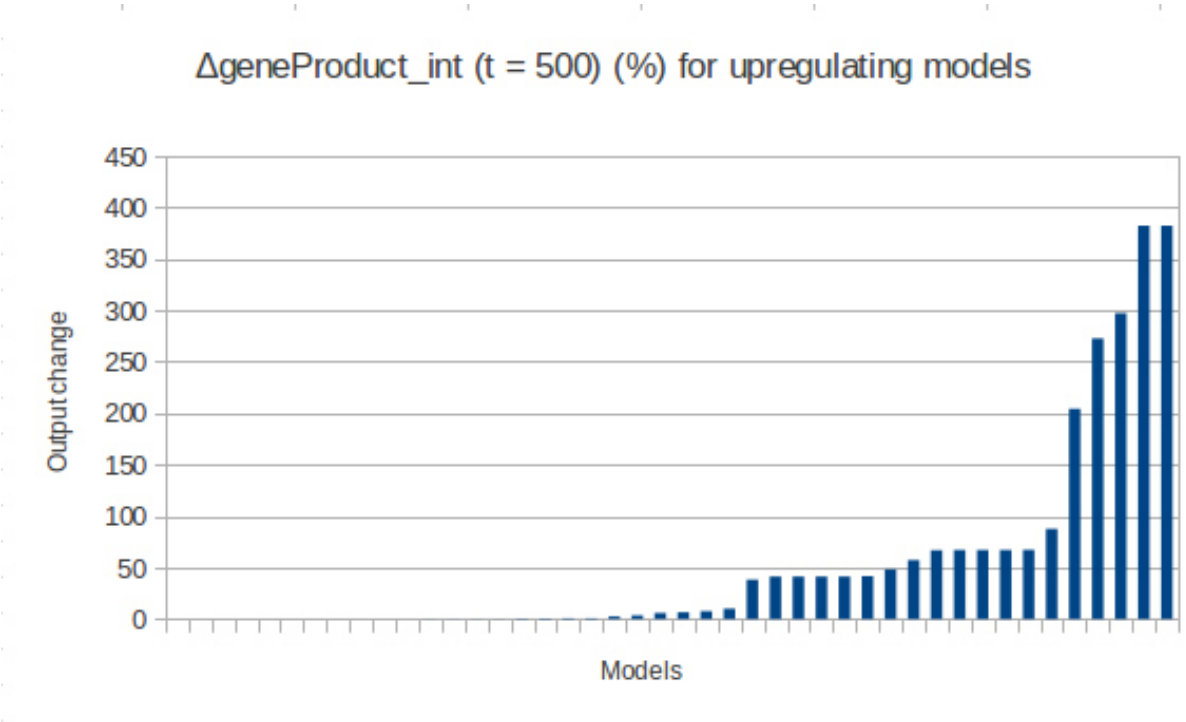

Figure 4.13: $\Delta geneProduct\_int_{(t=500)}$ **(%) for upregulated models** - This Figure shows the change of output for the set of upregulated mutant models, ordered from the lowest output to the highest. Figure has been generated with LibreOffice Calc.

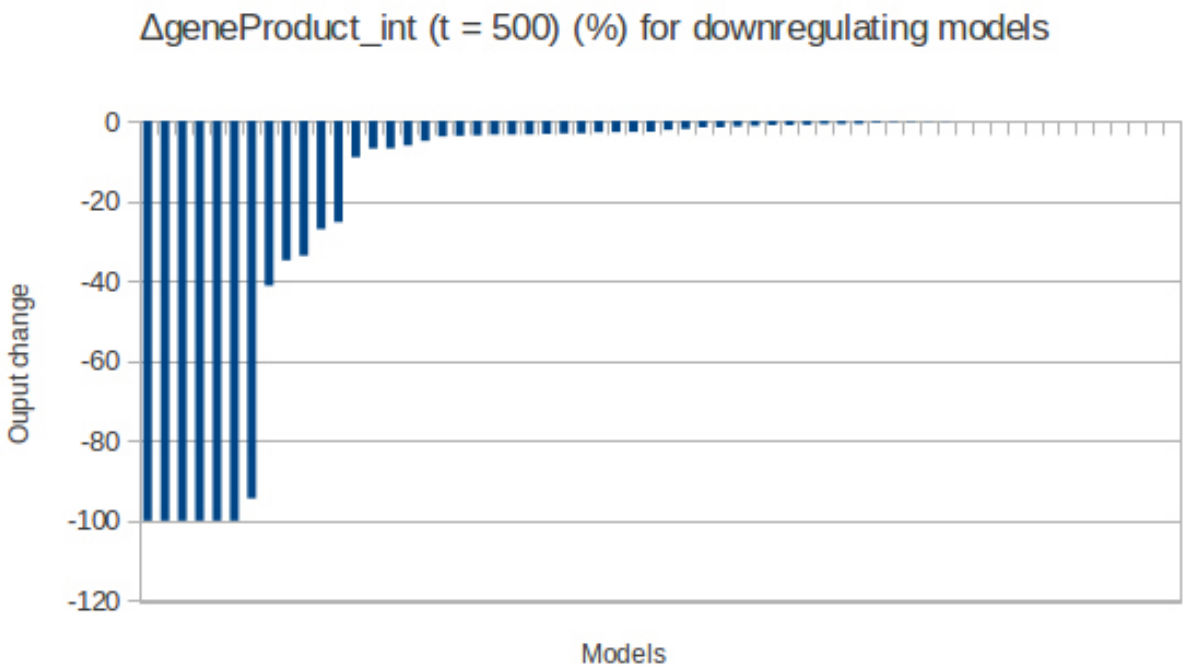

Figure 4.14: $\Delta geneProduct\_int_{(t=500)}$ (%) for downregulating models - This figure shows the change of output for the set of downregulated mutant models, ordered from the lowest output to the highest. Figure has been generated with LibreOffice Calc.

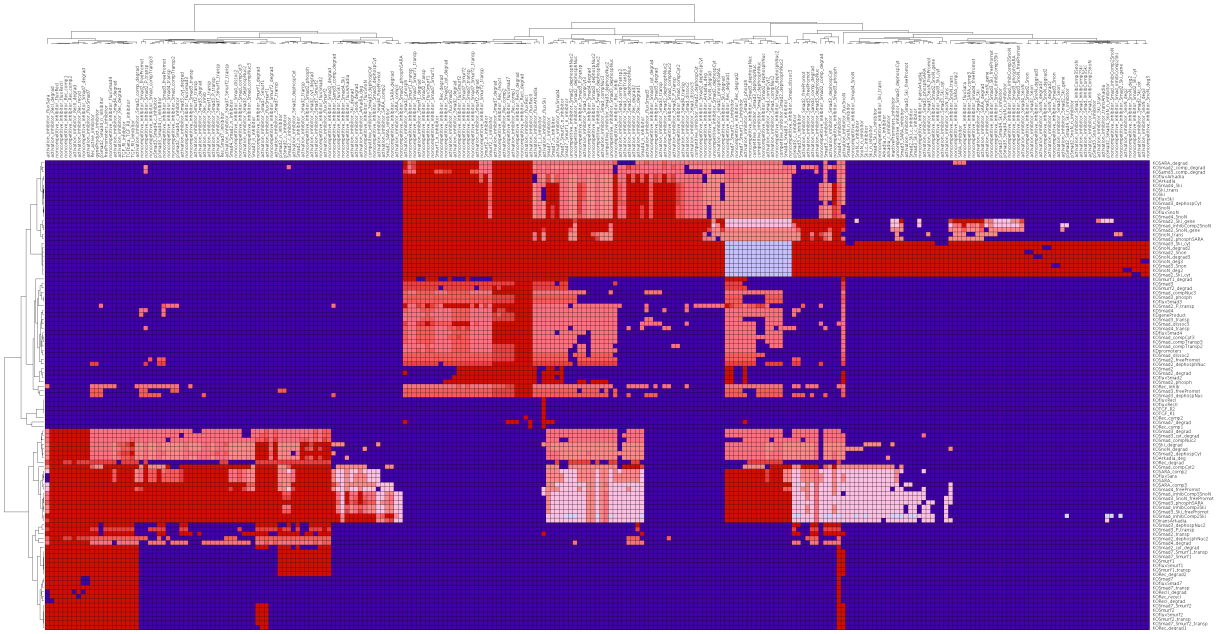

Figure 4.15: Same analysis as in Figure 4.6. -

-25%. To keep consistency with the analysis of the Zi model, I decided again to employ the least strict threshold. The results of the filtered analysis are shown in Figure 4.16 and summarised in Figure 4.17.

As it can be observed, drugs fall again into two groups, with the ones that restore upregulated models in the upper part and those that restore downregulated models in the lower part. upregulated models are, in general, restored by inactivators of the species essential for TGF-β signalling, such as the receptor complexes and $pRSmad\_Smad4\_n$, as well as by activating their degradation or inhibiting their synthesis. The possible array of drugs that restore the downregulated models is much smaller, and six of the most downregulated models can only be restored by a non-essential activator of Ski synthesis. The other models can, in general, be treated by targeting species implicated in negative feedback, i.e. $Smad7$ and $Smurf1/2$, or by drugs with opposite effect to the ones against upregulated models.

Especially interesting is the case of Ski. This species is part of the negative feedback mechanism of TGF-β signalling, as it binds to activated Smad heterodimers and blocks gene expression. Therefore, one would not expect an activator of its synthesis to restore downregulated models. However, when Ski binds to nuclear $Smad4$, it blocks $Smad7$ productions, and this effect seems to be relevant enough so that it can restore most of the downregulated models, even models with an output reduction near to 100%, such as $KORec\_comp1$ (see Figure **??**).

4.3 Biological insights gained from my framework

The two models on TGF-β signalling that I have analysed in this work focus on different aspects of the network. While the Zi model focuses on the basic aspects of signalling, which are shown in detail; the Wegner model is more complex as it describes the feedback mechanisms, but does not show the central pathway with the same level of detail. Their output is different as well. The first one only shows the accumulation of $PSmad2\_Smad4\_n$, which approaches a steady-state and the second one models the relation between $pRSmad\_Smad4\_n$ accumulation and gene expression, which shows an oscillatory behaviour. Therefore, the results obtained from both models are not identical, but comparable.

4.3.1 Potential drug targets in TGF-β signalling

Our results have shown that there are two main kinds of drugs, those that can restore upregulated models and those that can restore downregulated models. However, not all drugs proposed by

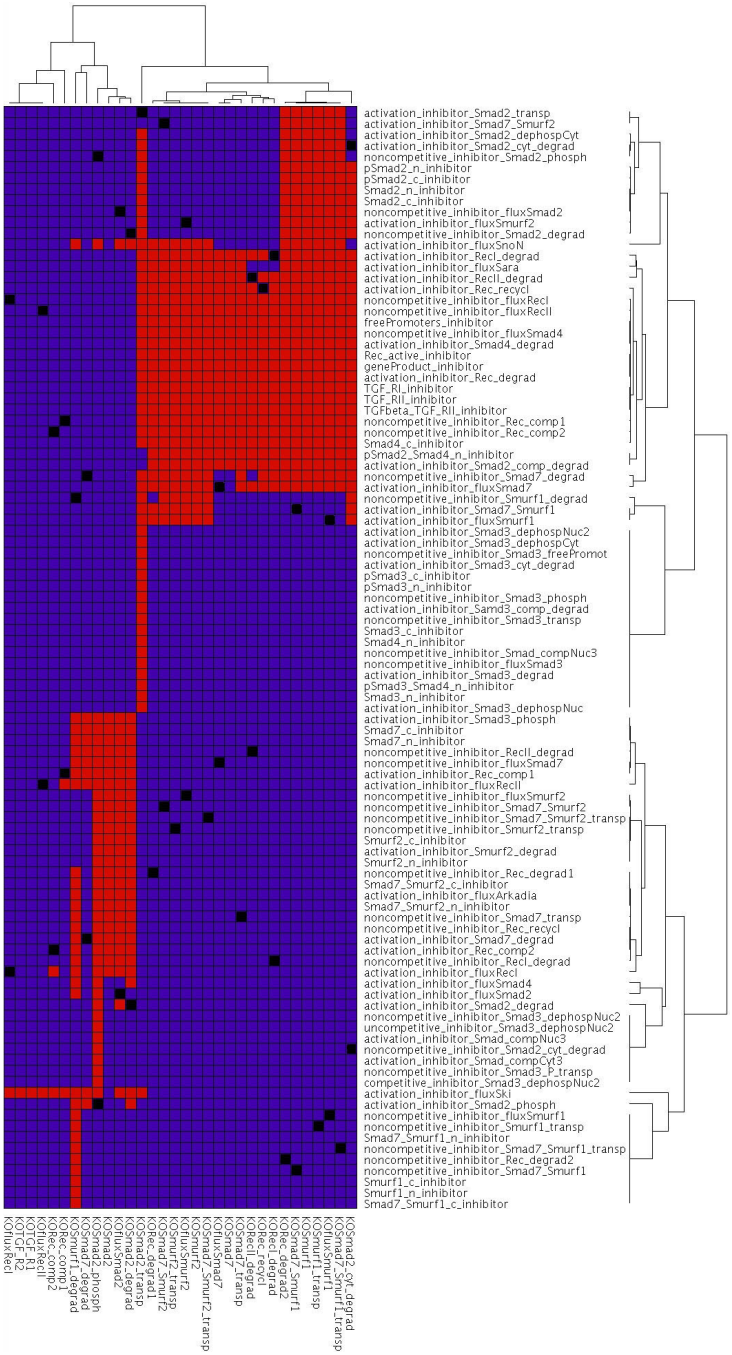

Figure 4.16: Successful treatments for diseased models of Wegner *et al* 2012 TGFβ model - Same analysis as in Figure 4.9.

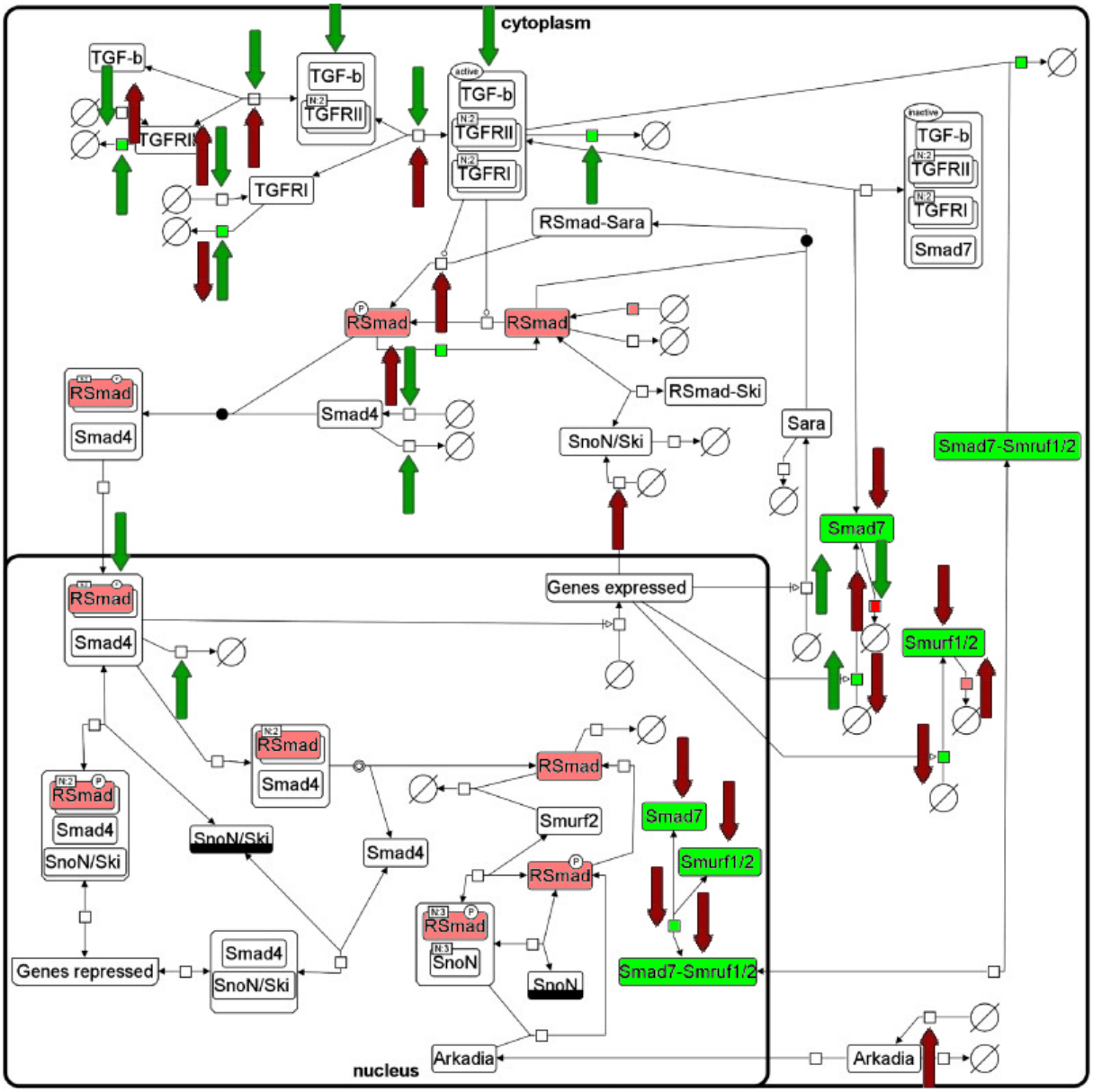

Figure 4.17: Same analysis as in Figure 4.3 -

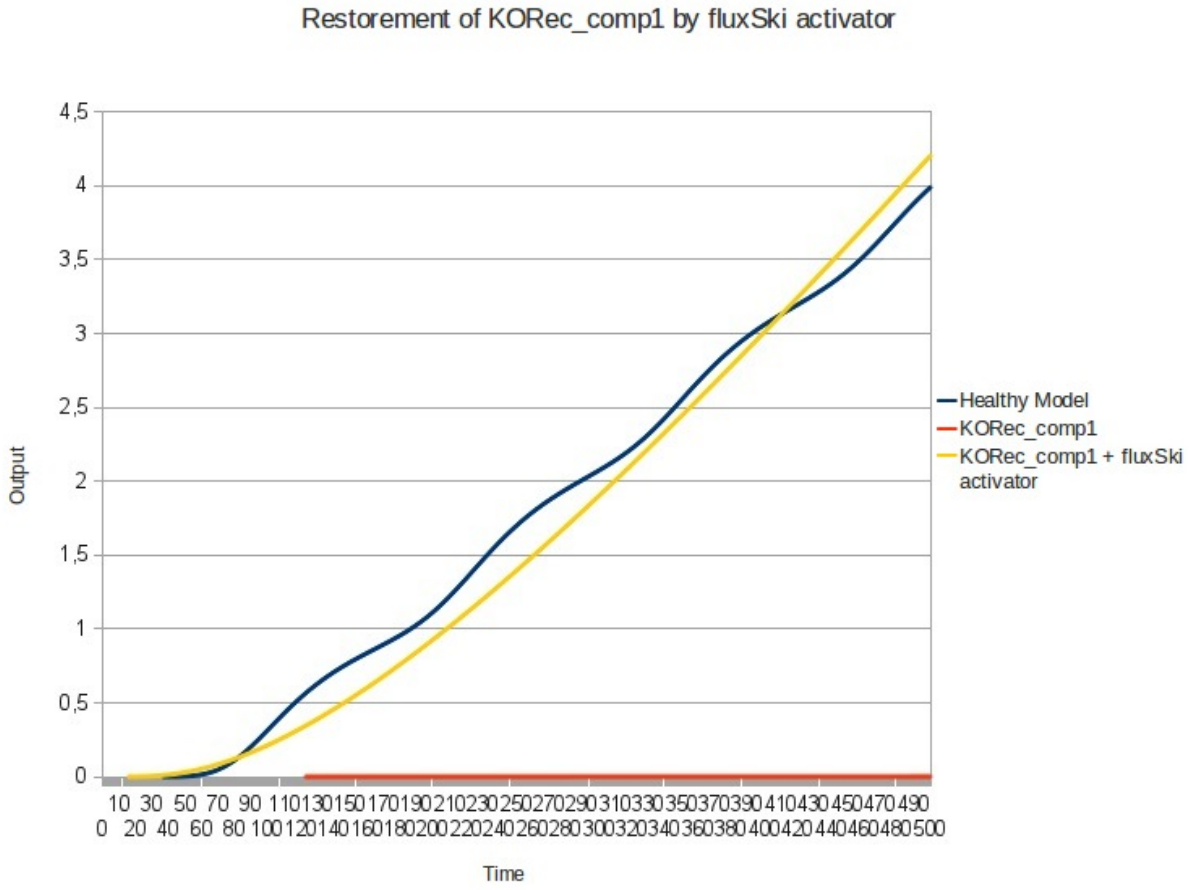

Figure 4.18: Restorement of the pathological model $KORec\_comp1$ by the non-essential activator of Ski synthesis - Time courses of the integral of $geneProduct$ versus time for the healthy model (blue), the pathological model $KORec\_comp1$ (red) and the same model, with a concentration of 7488.95 of $fluxSki$ activator, which satisfies that the objective function $f = 1$. Figure has been generated with LibreOffice Calc.

the tool are equally useful: non-essential activators and drugs targeting reactions that are not catalysed by an independent enzyme are generally more complicated to develop and thus, less valuable results. Therefore, I suggest a list of drug targets that I think are most interesting from a biological point of view:

- **Inhibitors of *RSmad* phosphorylation for diseases with upregulated TGF-β signalling**. The analysis on Zi model predict that inhibiting this reaction (reaction 18) will restore normal phenotype for diseases with downregulated TGF-β signalling.

- **Inactivators of Receptors, *RSmads* or *Smad*4 for diseases with upregulated TGF-β signalling**. The analysis on both models predicts that drugs targeting any of these molecules will be effective on diseases with upregulated it TGF-β signalling.

- **Inactivators of $Smurf2$ or *Smad*7 for diseases with TGF-β signalling downregulated** The analysis on Wegner model predicts that drugs targeting any of these molecules will be effective on diseases with downregulated TGF-β signalling.

- **Inactivators of $PPM1A$ for diseases with downregulated TGF-β signalling**. $PPM1A$ is a Ser/Thr protein phosphatase that dephosphorylates nuclear phosphorylated RSmad monomers(56). This reaction is modelled as reaction 24 in Zi model. The analysis predicts that an inhibitor of this reaction will be able to restore normal phenotype for diseases that downregulated TGF-β signalling.

The results provided by my framework are purely theoretical. Furthermore, they require certain assumptions to be accepted for the construction the objective function. These assumptions may deeply affect the results of the analysis. In this case, I have chosen to disregard all considerations on the shape of the response curve of TGF-β signalling, taking only one time point for the first model and the integral at a time point for the other model as reference. As I have observed that mutant models not only change the output of the system in a quantitative manner, but also affect the shape of the response curve, this could have an impact on the usefulness of certain drugs in human disease. However, as the employed models are not optimised for drug discovery for this analysis, I have decided to use less stringent conditions, maximising the amount of obtained output to avoid disregarding useful inhibitors that do not perform optimally for the mathematical model. As a next step, experiments should be conducted to check the proposed drug targets and confirm or reject the assumptions on which the analysis is based.

4.3.2 Experimental confirmation of the results

I have examined the existing literature for drugs targeting the specified components of the biochemical network. The only protein for which I found drugs targeting it was TGF-β Receptor I, which is specifically targeted by SB-431542, an inhibitor of its kinase activity. This drug, which is simulated as an inhibitor of reaction 18 on the Zi model, has been shown to be an effective drug against TGF-β dependent tumours in cell culture. This drug attenuates the tumour-promoting effects of TGF-β, including EMT, cell motility, migration and invasion(57). The fact that this drug is regarded as a potent antitumour agent(57) serves as a proof-of-concept for the validity of my framework.

To expand on this evidence, I propose an experimental approach for the confirmation of our results:

1. Creating a set of knock-downs, using iRNA technology, for genes related to the TGF-β signalling pathway.
2. Evaluation of signalling response curve with respect to time for every knock-down.
3. For every knock-down, a set of drugs is assayed and, if possible, dose-response at an arbitrary time is studied.
4. Signalling response curves with respect to time for every knock-down are studied when exposed to the optimal concentrations of drugs, and compared to healthy signalling response curve.

Knock-downs of the following genes are proposed for the assay:

- $PPM1A$. Predicted on Zi model to upregulate TGF-β signalling.
- $Smad7$. Predicted on Wegner model to upregulate TGF-β signalling.
- $Smurf1/2$. Predicted on Wegner model to upregulate TGF-β signalling.
- $Smad4$. Predicted on Zi model to downregulate TGF-β signalling.
- $TGF-\$\beta$ Receptor I (TR1).$PredictedonWegnermodeltodownregulateTGF-\beta$ signalling.

The following drugs are proposed for the assay, in agreement with the suggestions for drug targets presented in 4.3.1:

- SB-431542, inhibitor of $RSmad$ phosphorylation.
- Antibodies targeting TGF-β Receptor II (Anti-TR2), targeting TGF-β signalling.
- Interference RNA for $Smurf2$ and $Smad7$, targeting TGF-β feedback.
- Interference RNA for $PPM1A$.

The following assays are thus proposed:

- Treating $KDPPM1A$ with SB-431542.
- Treating $KDPPM1A$ with Anti-TR2.
- Treating $KDSmad7$ with SB-431542.
- Treating $KDSmad7$ with Anti-$TR2$.
- Treating $KDSmurf1/2$ with SB-431542.
- Treating $KDSmurf1/2$ with Anti-TR2.
- Treating $KDSmad4$ with $PPM1A$ iRNA.
- Treating $KDSmad4$ with $Smad7$ iRNA.
- Treating $KDSmad4$ with $Smurf2$ iRNA.
- Treating $KDTR1$ with $PPM1A$ iRNA.
- Treating $KDTR1$ with $Smad7$ iRNA.
- Treating $KDTR1$ with $Smurf2$ iRNA.

The performance of these assays would serve as an experimental confirmation or rejection for our findings in TGF-β signalling, which suggest the drug targets described in 4.3.1 are useful for treating various TGF-β signalling related diseases. Such results would serve as ultimate proof of the applicability of my framework in drug target discovery.

5

Discussion

5.1 What has been achieved

In this work I have presented a framework for the identification of effective drug targets against potential diseases on the basis of biochemical networks modelled by ordinary differential equations. The workflow of my framework comprises two steps, needing as only inputs an ODE model describing a healthy system and an objective function describing the main features that characterise the system as healthy. In the first step, the model is manipulated by systematically impairing every reaction and species, thus creating a set of models that simulate the action of possible mutations. In the second step, the models of this mutant set are subject to a drug scan, and different drug treatments with the ability of restoring the healthy behaviour of the system are determined. As a result, the tool provides the set of manipulated models and a list of the drugs that could restore the healthy behaviour for these models.

The results provided by this tool are purely theoretical, as they are only based on model structure and kinetics, not taking into account existing drugs, druggable proteins or protein annotations. It therefore produces a big quantity of output, which can then be interpreted based on general knowledge of the system. The results are consistent with biological evidence and provide enough information to hypothesise potential drug targets, as well as possible mutations that might affect the system. Moreover, one can obtain comparably fast and it does not require special resources. The analysis for the biggest model that I have studied here, lasts one to two days running on a single CPU, a very short time in comparison to the development of a drug. Because of all this, my tool can be potentially useful for helping to prioritise useful drug target candidates at the first step of drug development.

This framework expands on the work done on TIde (50). While TIde can be used for drug target selection in models that describe pathological conditions, my framework models these conditions automatically starting from a model of the healthy state. This provides a conceptual advantage for drug target search in multifactorial diseases, including disorders of the relevance of cancer or diabetes, which have a great impact in human health. In these multifactorial diseases, different mutations can cause similar phenotypes, so that only one model cannot account for the possible alterations in the system. In my framework, I overcome this difficulty by analysing all possible mutations in the healthy system, thus offering a list of drugs that are effective for a great number of mutants.

Moreover, my framework shows essential differences with respect to other approaches for drug target selection, which can be used to obtain complementary information on the biological systems. It provides a different output in comparison to other approaches, including stoichiometric approaches, flux analysis approaches and structural approaches. Stoichiometric approaches (58) identify choke points on the network, determining which reactions are essential for the synthesis and degradation of all metabolites, which are thus interesting for drug targeting. However, as they do not incorporate kinetic information, they cannot calculate which molecules are more sensitive to drug targeting. These approaches therefore provide a big quantity of output, but cannot evaluate targets for their efficacy and thus, as unable to prioritise more efficacious targets. My framework however, as it incorporates kinetic information, is able to disregard those drug targets with a weak effect on the network, and prioritise working drug targets according to the minimal concentration needed for model restoration.

Flux balance analysis-based approaches (59) evaluate the influence of drugs on the fluxes through the network. However, they cannot study the dynamic response of the system to the drug, and do not offer quantitative results of species concentrations, which my tool does. As the evaluation of fluxes instead of concentrations can lead to different results for the same model (51), these approaches can be used for the confirmation of the biological relevance of the results obtained by my tool.

Lastly, structural approaches integrate protein structural and sequence information into target prediction, which can be useful for determining which proteins are most likely druggable. These approaches, however, do not study the dynamic response of the system or prioritise drug targets according to their efficacy, which my tool does. Further improvements on my framework would include the integration of these sources of information, allowing for the evaluation of fluxes within biochemical networks, the incorporation of structural information of the components of

the route and the determination of their druggability, including the information provided by drug databases, such as Drugbank (60) and STITCH 2 (61).

5.2 Application of the framework to biological problems

The study on TGFβ signalling pathway provides an example of the application of my tool to a real biological problem, showing its usefulness for providing a list of potential drug targets with the capacity of modulating the activity of a network. However, there are important issues that remain to be addressed. First and foremost, there are numerical problems with some inhibitors which produce spurious results. These problems can be manually detected and solved by changing the tolerance parameters of the numerical solver. However, generally detecting and handling this problem by adapting the solver options to the model is beyond the scope of this work.

The second issue is determining the most adequate definition for the objective function which describes the healthy state of the model. In this work I have used an open criterion for the definition of the objective function, basing it on the value of one species at one time point instead of describing the whole response curve from ($t = 0$). This has been done in order to maximise the obtaining of output, and avoid eliminating potentially useful targets. However, other stricter criteria are possible, as the tool allows defining the objective function for multiple species and time points. Whether stricter or more open approaches are better suited for the retrieval of biologically relevant results is an issue that can only be addressed by experimental validation of the results.

A third issue that needs to be addressed is determining which models can be analysed by this framework. Models are created with the aim of describing certain characteristics of the studied system. None of the models I have employed for demonstrating its applicability have been specifically designed with the objective of drug discovery, yet I have been able to obtain results with a possible relevance to drug discovery. In general, this depends on the level of mechanistic detail of the model, as a more mechanistic model will have species and reactions that closely correspond to real proteins or reactions catalysed by proteins, so that the set of mutant models will correspond to possible mutations in the genes that code them. For example, for the case of TGF-β signalling, the study of feedback mechanisms is much more useful if these mechanisms are individually modelled and not as a general reaction. Therefore, in this case, the

results obtained from the Wegner model for drug targeting on the feedback mechanisms might be of higher relevance than for the Zi model.

Lastly, the impact of uncertainties in the models needs to be assessed. Most of the parameters in a model are not experimentally determined, but through a process of parameter estimation which fits the model into the experimental data. The process has a certain degree of uncertainty, since it is possible that more than one parameter set provides equally good model fittings. Since the probability of non-identifiability increases with higher model complexity, it especially affects very detailed models, which are the most appropriate subject for my framework. Uncertainty might thus have an impact on the quality of the results obtained by my framework, as the biological relevance of the parameters cannot be assured. Therefore, to assess the importance of these uncertainties, I propose applying the framework for a set of models accounting for a range of possible parameters set, in order to detect which results appear consistently for all parameters sets. This approach has been implemented with different variations before (51, 62**? ?**). As we do not have a reason to favour one parameter set over the others, this can serve for prioritising drug targets which are effective at a wider range of conditions.

5.3 The cycle of systems biology and drug discovery

The final objective of my framework is making a set of predictions about which molecules in the system should be targeted in order to rescue a potential disease. These predictions have no backing experimental evidence behind them, and the model on which predictions are based might not be optimised for drug discovery. Therefore, additional experiments are needed to check the feasibility of my framework's predictions. The model and the objective function thus enter the cycle of systems biology, the cornerstone of computational modelling and systems biology. In the cycle, the predictions generated by the model and its objective function are experimentally tested. Given that the results confirm the initial predictions, the model will be considered a useful tool for drug target prediction and the objective function an accurate description of the main characteristics of healthy state. However, if the experiments do not support the predictions, either the model, the objective function, or both, have to be modified so that the predictions they generate are in accordance with new evidence. In the particular case of TGF-β, the testing of the hypothesis generated in Section 4.3.2 will shed light on the fitness of the models employed for drug scanning, as well as evaluate whether an open or a strict criterion should be used for classifying a system as healthy.

We are still at the very first steps of applying systems biology to drug discovery. Computational models are not trustworthy enough to become the main tools for drug target selection. Drug scan tools have provided interesting results (see (27)) but, as far as I know, no successful drug has been developed using these tools. For these first steps, computational modelling might only have a secondary role for drug target selection, in comparison to the omic technologies. But systems biology has a cyclic nature: models and approaches will be refined by incorporating more and more experimental data, becoming more reliable and informative. Therefore, biological knowledge will be incorporated in an integrated way, with the final objective of understanding diseases and how to cure them.

6

Appendix 1

Source code of the tool:

```
#!/usr/bin/env python

import sys, re, copy
import libTI2.global_options
import libTI2.model
import libTI2.laotide
import libTI2.solvers.lsodasolver
import math
import scipy.stats
import libsbml
from optparse import OptionParser

def addtime(model):
    model.parameters.append("time")
    model.parameter_values["time"] = 0.0
    model.differential_equations["time"] = " 0 + 1.0"

def load_timecourses_to_fit_to(timecoursefilename):
    f = open(timecoursefilename,"r")
    lines = f.readlines()
    f.close()
    delimiter = ","
    time_key = "t"
    allkeys = lines[0].strip("\n").split(delimiter)
    time_key_index = allkeys.index(time_key)
    substance_keys = allkeys[:time_key_index]+allkeys[time_key_index+1:]
```

```
27      timecourses_to_fit_to = {}
28      for line in lines[1:]:
29          if len(line.strip("\n"))==0 or line.startswith("#"):
30              continue
31          vals = []
32          for x in line.strip("\n").split(delimiter):
33              if x=="":
34                  vals.append(None)
35              else:
36                  if "[" in x and "]" in x:
37                      entry = x.strip(" []").split(";")
38                      vals.append([float(x) for x in entry])
39                  else:
40                      vals.append(float(x))
41          timecourses_to_fit_to[vals[time_key_index]] = dict(zip(
              substance_keys,vals[:time_key_index]+vals[time_key_index+1:]))
42      return timecourses_to_fit_to
43  def calculate_timepoint(timecourses):
44      sum_time = 0.0
45      count = 0
46      for timepoint in timecourses:
47          sum_time += timepoint
48          count += 1
49      timepoint = sum_time/count
50      return timepoint
51  def integral_timecourse(timecourses, model):
52      int_species = []
53      for timepoint in timecourses:
54          new_dict = {}
55          for expression in timecourses[timepoint]:
56              for species in model.species:
57                  if species in expression:
58                      new_expression = expression.replace(species, species+"
                            _int")
59                      int_species+=[species]
60              new_dict[new_expression] = timecourses[timepoint][expression]
61          timecourses[timepoint] = new_dict
62      return int_species
63
64
```

```
65 def doc_analysis(doc_filename, definitions_dict, model, inhibitor_groups
       =False):
66     doc_file = open(doc_filename, "r")
67     doc_list = doc_file.readlines()
68     doc_file.close()
69     for line in doc_list:
70         if ":" not in line:
71             continue
72         definition = line.split(": ")[0]
73         definitions_list = line.split(": ")[1].strip("\n").split(", ")
74         if definitions_list == []:
75               print "ERROR: Missing Species or Reaction on line number "+
                      str(doc_list.index(line)+1)+" of "+ doc_filename.split("/
                      ")[-1]+".\n This line will be ignored."
76         elif definition in model.species or definition in model.
               inhibited_reaction_equations:
77             print "ERROR: Species and Reaction Groups cannot have the name
                   of an existing species. Line "+str(doc_list.index(line)+1)+"
                    of "+ doc_filename.split("/")[-1]+" will be ignored."
78
79         else:
80             is_species = False
81             is_reaction = False
82             not_exists = False
83             for form in definitions_list:
84                 if form in model.species:
85                     is_species = True
86                 elif form in model.inhibited_reaction_equations:
87                     is_reaction = True
88                 else:
89                     not_exists = True
90             if not_exists:
91                 print "ERROR: A value on line number "+str(doc_list.index(
                        line)+1)+" of "+ doc_filename.split("/")[-1]+" does not
                        correspond to any Species or Reaction in the model.\n
                        This line will be ignored."
92             elif is_species and is_reaction:
93                 if inhibitor_groups:
94                     print "ERROR: Inhibitor Groups cannot target both species
                           and reactions. Therefore, line "+str(doc_list.index(
                          line)+1)+" of "+ doc_filename.split("/")[-1]+" will be
```

```
                    ignored."
95              else:
96                  print "ERROR: Correspondences file cannot contain both
                        species and reactions in the same line. Therefore,
                        line "+str(doc_list.index(line)+1)+" of "+
                        doc_filename.split("/")[-1]+" will be ignored."
97          elif is_species:
98              if inhibitor_groups:
99                  definitions_dict[definition+"_inhibitor"] = [form + "
                        _inhibitor" for form in definitions_list]
100             else:
101                 definitions_dict[definition] = definitions_list
102         elif is_reaction:
103             if inhibitor_groups:
104                 definitions_dict["activation_inhibitor_"+definition] = ["
                        activation_inhibitor_"+form for form in
                        definitions_list]
105                 definitions_dict["noncompetitive_inhibitor_"+definition]
                        = ["noncompetitive_inhibitor_"+form for form in
                        definitions_list]
106             else:
107                 definitions_dict[definition] = definitions_list
108
109 def add_other_components(model, definitions_dict, inhibitor_groups =
        False):
110     if inhibitor_groups:
111         array = [model.inhibitor_names]
112     else:
113         array = [model.species, model.inhibited_reaction_equations]
114     for componentlist in array:
115         for component in componentlist:
116             is_already_in_dict = False
117             for form in definitions_dict:
118                 if component in definitions_dict[form]:
119                     is_already_in_dict = True
120                     break
121             if not is_already_in_dict:
122                 definitions_dict[component] = [component]
123
124 def addinhibitors(model, sbmlmodel):
125     newinhibitors = []
```

```
126     inactives = []
127     compartment_dict = {}
128     species_volume = {}
129     readmodel = sbmlmodel.getModel()
130     for compartment in readmodel.getListOfCompartments():
131         compartment_dict[compartment.getId()] = str(compartment.getVolume
            ())
132     for species in readmodel.getListOfSpecies():
133         species_name = species.getId()
134         if species_name in model.species:
135             if species.isSetInitialConcentration():
136                 volume = "1.0"
137             else:
138                 volume = compartment_dict[species.getCompartment()]
139             inhibitor = species_name+"_inhibitor"
140             inactive = species_name+"_inactive"
141             adddifferentialeq(model, species_name, inhibitor, inactive,
                volume)
142             addtolists(model, inhibitor, inactive)
143 
144 def adddifferentialeq(model, species, inhibitor, inactive, volume):
145     model.new_differential_equations[species] += " - 1.0 * ( "+species+"
        / "+volume+" ) * ( "+inhibitor+" ) + 1.0 * ( "+inactive+" / "+
        volume+" )"
146     model.new_differential_equations[inactive] = " 1.0 * ( "+species+" /
        "+volume+" ) * ( "+inhibitor+" ) - 1.0 * ( "+inactive+" / "+volume
        +" )"
147 
148 def addtolists(model, inhibitor, inactive):
149     model.new_species.append(inhibitor)
150     model.inhibitor_names.append(inhibitor)
151     model.species.append(inactive)
152     model.new_species_values[inhibitor] = 0.0
153     model.species_values[inactive] = 0.0
154     model.new_species_values[inactive] = 0.0
155 
156 def defineintervals(minimal, maximal, stepnumber):
157     step = (math.log(maximal)-math.log(minimal))/(stepnumber-1)
158     concentrations = [minimal]
159     for i in range(stepnumber-1):
160         concentrations += [math.exp(math.log(concentrations[i])+step)]
```

```
161     return concentrations
162 
163 def addintegral(model, species):
164     model.species.append(species+"_int")
165     model.species_values[species+"_int"] = 0.0
166     model.new_differential_equations[species+"_int"] = " 0 + 1.0 * ( "+
            species+" )"
167 
168 def knockout(model, corr_dict, species):
169     model.model_name = "KO"+species
170     for form in corr_dict[species]:
171         if form in model.species:
172             kospecies(model, form)
173         if form in model.inhibited_reaction_equations:
174             koreaction(model, form)
175 
176 def kospecies(model, species):
177     inhibitor = species+"_inhibitor"
178     model.new_species_values[inhibitor] = 100000000.0
179 
180 def koreaction(model, reaction):
181     inhibitor = "noncompetitive_inhibitor_"+reaction
182     model.new_species_values[inhibitor] = 99999999.0
183 
184 def knockdown(model, corr_dict, species):
185     model.model_name = "KD"+species
186     times = {}
187     for form in corr_dict[species]:
188         if form in model.species:
189             kdspecies(model, form, times)
190         if form in model.inhibited_reaction_equations:
191             kdreaction(model, form)
192 
193 def kdspecies(model, species, times):
194     inhibitor =  species+"_inhibitor"
195     if species not in times:
196         times[species] = 1
197     inhibition = float(2**times[species] - 1)
198     model.new_species_values[inhibitor] = inhibition
199     times[species] += 1
200 
```

```
201 def kdreaction(model, reaction):
202     inhibitor = "noncompetitive_inhibitor_"+reaction
203     if model.new_species_values[inhibitor] == 0.0:
204         model.new_species_values[inhibitor] = 1.0
205 
206 def restore(model, corr_dict, species):
207     for form in corr_dict[species]:
208         if form in model.species:
209             inhibitor = form+"_inhibitor"
210             model.new_species_values[inhibitor] = 0.0
211         if form in model.inhibited_reaction_equations:
212             inhibitor = "noncompetitive_inhibitor_"+form
213             model.new_species_values[inhibitor] = 0.0
214 
215 def analysis_loop(model, solver1, solver2, inhibitor_dict, chicut_value,
        corr_dict, value, concentrationsTIde, outputfile):
216     restorable = False
217     print model.model_name
218     outputfile.write(model.model_name+"\n")
219     constantx = [model.species_values[s] for s in model.species]
220     x = [model.new_species_values[s] for s in model.inhibitor_names]
221     solver1.solve(starting_species_concentrations=(constantx+x))
222     outputfile.write(str(solver1.rssq)+"\n")
223     root = math.sqrt(solver1.rssq)
224     percentage = (root - 1.0)*100.0
225     print str(percentage)+"%"
226     outputfile.write(str(percentage)+"\n")
227     if solver1.rssq > 1.0:
228         solver = solver1
229         print "Upregulated model"
230         outputfile.write("Upregulated model\n")
231     if solver1.rssq == 1.0:
232         print "Normal model"
233         outputfile.write("Normal model\n")
234         return True
235     if solver1.rssq < 1.0:
236         solver = solver2
237         print "Downregulated model"
238         outputfile.write("Downregulated model\n")
239     for inhibitorgroup in inhibitor_dict:
```

```
240         effective = one_dimensional_analysis(solver, model, inhibitor_dict
                , inhibitorgroup, chicut_value, corr_dict, value,
                concentrationsTIde, outputfile)
241         if effective is None:
242             break
243         if effective:
244             restorable = True
245     outputfile.write("\n")
246     return restorable
247
248 def one_dimensional_analysis(solver, model, inhibitor_dict,
    inhibitorgroup, chicut_value, corr_dict, value, concentrationsTIde,
    outputfile):
249     i = 0
250     two_values_left = {}
251     two_values_right = {}
252     successful = False
253     ko_in_group = False
254     for inhibitor in inhibitor_dict[inhibitorgroup]:
255         if tidemodel.new_species_values[inhibitor] != 0.0:
256             ko_in_group = True
257         elif "_inhibitor_" in inhibitor:
258             index = inhibitor.find("_inhibitor_")
259             reaction = inhibitor[11+index:]
260             if reaction in corr_dict[value]:
261                 ko_in_group = True
262     if ko_in_group:
263         return successful
264     while i < len(concentrationsTIde):
265         for inhibitor in inhibitor_dict[inhibitorgroup]:
266             model.new_species_values[inhibitor] = concentrationsTIde[i]
267         constantx = [model.species_values[s] for s in model.species]
268         x = [tidemodel.new_species_values[s] for s in tidemodel.
                inhibitor_names]
269         solver.solve(starting_species_concentrations=(constantx+x))
270         if solver.rssq is None:
271             for inhibitor in inhibitor_dict[inhibitorgroup]:
272                 model.new_species_values[inhibitor] = 0.0
273             return solver.rssq
274         if solver.rssq > chicut_value:
275             if concentrationsTIde[i-1] not in two_values_left and i!= 0:
```

```
276             two_values_left = {}
277         elif len(two_values_left) == 2 and concentrationsTIde[i-2] in
                two_values_left:
278             del two_values_left[concentrationsTIde[i-2]]
279         two_values_left[concentrationsTIde[i]] = solver.rssq
280     if solver.rssq < chicut_value:
281         if concentrationsTIde[i-1] not in two_values_right:
282             two_values_right = {}
283         two_values_right[concentrationsTIde[i]] = solver.rssq
284 
285         if len(two_values_right) == 2 and len(two_values_left) > 0:
286             successful = True
287             four_values_dict = dict(two_values_left.items() +
                    two_values_right.items())
288             chicut, std_err = calculatechicut(four_values_dict,
                    chicut_value)
289             if chicut < concentrationsTIde[0]:
290                 print inhibitorgroup +" is effective for restoring " +
                        model.model_name + " for a concentration lower than "+
                        str(concentrationsTIde[0])
291                 outputfile.write(inhibitorgroup +" is effective for
                        restoring " + model.model_name + " for a concentration
                         lower than "+str(concentrationsTIde[0])+"\n")
292             else:
293                 print inhibitorgroup +"("+str(chicut)+"): is effective
                        for restoring " + model.model_name
294                 outputfile.write(inhibitorgroup +"("+str(chicut)+"): is
                        effective for restoring " + model.model_name + "\n")
295             break
296 
297         elif len(two_values_right) == 2:
298             successful = True
299             print inhibitorgroup +" is effective for restoring " + model
                    .model_name + " for a concentration lower than "+str(
                    concentrationsTIde[0])
300             outputfile.write(inhibitorgroup +" is effective for
                    restoring " + model.model_name + " for a concentration
                    lower than "+str(concentrationsTIde[0])+"\n")
301             break
302     i += 1
303 for inhibitor in inhibitor_dict[inhibitorgroup]:
```

```
304         model.new_species_values[inhibitor] = 0.0
305     return successful
306 
307 def calculatechicut(four_values_dict, chicut_value):
308     x = [math.log(s) for s in four_values_dict]
309     y = [four_values_dict[s] for s in four_values_dict]
310     gradient, intercept, r_value, p_value, std_err = scipy.stats.
            linregress(x,y)
311     chicutlog = (1 - intercept)/gradient
312 
313     chicut = math.exp(chicutlog)
314     return chicut, std_err
315 
316 def summarize(content, summary_filename):
317     summary = open(summary_filename, "w")
318     counter = 0
319     while counter < len(content):
320         a = re.match("\d", content[counter])
321         if a:
322             chi = float(content[counter])
323             root = math.sqrt(chi)
324             percentage = (root - 1.0)*100.0
325             i = 0
326             if math.fabs(percentage) >= 10.0:
327                 summary.write(content[counter-1])
328                 summary.write(str(root)+"\n"+str(percentage)+"%\n")
329 
330                 while 1:
331                     result = re.match("\w", content[counter+i+2])
332                     if not result:
333                         break
334                     else:
335                         i += 1
336                 if i == 1:
337                     summary.write("No drug can restore model\n")
338                 else:
339                     summary.write(str(i-1)+" drugs can restore model\n")
340             counter += i + 2
341         counter += 1
342 usage = "usage: %prog [options] input_file output_directory
        time_course_1 time_course_2"
```

```
343 parser = OptionParser(usage=usage)
344
345 parser.add_option("-i", "--int",
346                   action="store_true", dest="int", default=False,
347                   help="the time courses refer to the integrals of the
                          variables (default: False)")
348 parser.add_option("-c", "--corr",
349                   action="store", dest="corr", default=False,
350                   help="directory of the file describing the reactions
                          and species that are simultaneously knocked-out")
351 parser.add_option("-d", "--coinh",
352                   action="store", dest="inh", default=False,
353                   help="directory of the file describing the reactions
                          and species that are coinhibited")
354 parser.add_option("-l", "--low",
355                   action="store", type="float", dest="low", default
                          =0.01,
356                   help="lowest drug concentration employed in the
                          optimization (default: 0.01)")
357 parser.add_option("-j", "--high",
358                   action="store", type="float", dest="high", default
                          =10000.,
359                   help="highest drug concentration employed in the
                          optimization (default: 10000.0)")
360 parser.add_option("-s", "--step",
361                   action="store", type="int", dest="step", default=50,
362                   help="number of steps employed in the optimization (
                          default: 50)")
363
364 (options, args) = parser.parse_args()
365
366 if len(args) != 4:
367    print "Incorrect input command"
368    sys.exit()
369 sbml_filename = args[0]
370 output_path = args[1]
371 timecourse_filename1 = args[2]
372 timecourse_filename2 = args[3]
373
374
375 chicut_value = 1
```

```
corr_dict = {}
inhibitor_dict={}
reactions_list=[]

model_name = sbml_filename.split("/")[-1][:-4]
go = libTI2.global_options.global_options()
go.initialize()

savedmodel = libTI2.model.model.Model(sbml_filename)
sbmlmodel = libsbml.readSBML(sbml_filename)

addtime(savedmodel)

tidemodel =  libTI2.laotide.laotide_model(savedmodel, replace_parameters
   =True, insert_inhibitions=True)

timecourses1 = load_timecourses_to_fit_to(timecourse_filename1)
timecourses2 = load_timecourses_to_fit_to(timecourse_filename2)

timepoint = calculate_timepoint(timecourses1)

del tidemodel.model_instance.sbmldoc
del tidemodel.model_instance.aih

if options.corr:
   doc_analysis(options.corr, corr_dict, tidemodel)

if options.coinh:
   doc_analysis(options.inh, inhibitor_dict, tidemodel, inhibitor_groups
       = True)

add_other_components(tidemodel, corr_dict)

addinhibitors(tidemodel, sbmlmodel)

add_other_components(tidemodel, inhibitor_dict, inhibitor_groups = True)


concentrationsTIde = defineintervals(options.low, options.high, options.
   step)

```

```
414 if options.int:
415     int_species1 = integral_timecourse(timecourses1, tidemodel)
416     int_species2 = integral_timecourse(timecourses2, tidemodel)
417     int_species = set(int_species1+int_species2)
418     for species in int_species:
419         addintegral(tidemodel, species)
420 
421 solver1 = libTI2.solvers.lsodasolver.lsodasolver(tidemodel.species,
422                     tidemodel.species_values,
423                     tidemodel.new_differential_equations,
424                     False,
425                     parameters=tidemodel.inhibitor_names,
426                     timecourses_to_fit_to=timecourses1)
427 #                   additional_options={"rtol":"1.0d-15","atol":"1.0d
        -18"})
428 
429 solver2 = libTI2.solvers.lsodasolver.lsodasolver(tidemodel.species,
430                     tidemodel.species_values,
431                     tidemodel.new_differential_equations,
432                     False,
433                     parameters=tidemodel.inhibitor_names,
434                     timecourses_to_fit_to=timecourses2)
435 #                   additional_options={"rtol":"1.0d-15","atol":"1.0d
        -18"})
436 
437 
438 if output_path[-1] != "/":
439     output_path += "/"
440 
441 outputfilename = output_path + model_name + "_Output.txt"
442 summary_filename = output_path + model_name + "_Summary.txt"
443 
444 outputfile = open(outputfilename, "w")
445 outputfile.write("TIde Analysis Output for Model "+model_name+"\n\n")
446 constantx = [tidemodel.species_values[s] for s in tidemodel.species]
447 x = [tidemodel.new_species_values[s] for s in tidemodel.inhibitor_names]
448 solver1.solve(starting_species_concentrations=(constantx+x))
449 for value in corr_dict:
450     knockout (tidemodel, corr_dict, value)
451     restorable = analysis_loop(tidemodel, solver1, solver2,
            inhibitor_dict, chicut_value, corr_dict, value, concentrationsTIde
```

```
        , outputfile)
452     if not restorable:
453         print "No drug can restore "+tidemodel.model_name
454         outputfile.write("No drug can restore "+tidemodel.model_name+"\n\n
                ")
455         restore (tidemodel, corr_dict, value)
456         knockdown (tidemodel, corr_dict, value)
457         restorable = analysis_loop(tidemodel, solver1, solver2,
                inhibitor_dict, chicut_value, corr_dict, value,
                concentrationsTIde, outputfile)
458         if not restorable:
459             print "No drug can restore "+tidemodel.model_name
460             outputfile.write("No drug can restore "+tidemodel.model_name+"\
                    n\n")
461     restore(tidemodel, corr_dict, value)
462 outputfile.close()
463 read_outputfile = open(outputfilename, "r")
464 content = read_outputfile.readlines()
465 read_outputfile.close()
466 summarize(content, summary_filename)
```

7

Appendix 2

Correspondences file used for the analysis of the Zi *et al* 2011 TGFβ model, indicating the species that are knocked-out together (right) under a common name (left):

```
1 TGF_beta: TGF_beta_ex, TGF_beta_ns, TGF_beta_endo
2 T1R: T1R_surf, T1R_endo, LRC_surf, LRC_endo
3 T2R: T2R_surf, T2R_endo, LRC_surf, LRC_endo
4 Smad2: Smad2c, PSmad2c, Smad2n, PSmad2n, PSmad2_PSmad2_c,
    PSmad2_PSmad2_c, PSmad2_PSmad2_n, PSmad2_PSmad2_n, PSmad2_Smad4_c,
    PSmad2_Smad4_n
5 Smad4: Smad4c, Smad4n, PSmad2_Smad4_c, PSmad2_Smad4_n
```

Correspondences file used for the analysis of the Wegner *et al* 2012 TGFβ mode:

```
1 Arkadia: Arkadia_c, Arkadia_n
2 promoters: freePromoters, inactivePromoters
3 Smad2: pSmad2_c, pSmad2_n, pSmad2_Smad4_c, pSmad2_Smad4_c,
    pSmad2_Smad4_n, pSmad2_Smad4_n, pSmad2_Smad4_Ski_n,
    pSmad2_Smad4_Ski_n, pSmad2_Smad4_SnoN_n, pSmad2_Smad4_SnoN_n,
    pSmad2_SnoN_n, pSmad2_SnoN_n, pSmad2_SnoN_n, Smad2_c, Smad2_n,
    Smad2_SARA, Smad2_Ski_c, Smad4_Smad2_n
4 Smad4: pSmad2_Smad4_c, pSmad2_Smad4_n, pSmad2_Smad4_Ski_n,
    pSmad2_Smad4_SnoN_n, pSmad3_Smad4_c, pSmad3_Smad4_n,
    pSmad3_Smad4_Ski_n, pSmad3_Smad4_SnoN_n, Smad4_c, Smad4_n,
    Smad4_Ski_n, Smad4_Smad2_n, Smad4_Smad3_n, Smad4_SnoN_n
5 Smad3: pSmad3_c, pSmad3_n, pSmad3_Smad4_c, pSmad3_Smad4_c,
    pSmad3_Smad4_n, pSmad3_Smad4_n, pSmad3_Smad4_Ski_n,
    pSmad3_Smad4_Ski_n, pSmad3_Smad4_SnoN_n, pSmad3_Smad4_SnoN_n,
    pSmad3_SnoN_n, pSmad3_SnoN_n, pSmad3_SnoN_n, Smad3_c, Smad3_n,
    Smad3_SARA, Smad3_Ski_c, Smad4_Smad3_n
```

```
6  Smad7: Rec_Smad7, Smad7_c, Smad7_n, Smad7_Smurf1_c, Smad7_Smurf1_n,
       Smad7_Smurf2_c, Smad7_Smurf2_n
7  TGF_R1: TGF_RI, Rec_active, Rec_active, Rec_Smad7, Rec_Smad7
8  TGF_R2: TGF_RII, Rec_active, Rec_Smad7, TGFbeta_TGF_RII, TGFbeta_TGF_RII
9  SARA_: SARA, Smad3_SARA, Smad2_SARA
10 Ski: Ski_c, Ski_n, pSmad2_Smad4_Ski_n, pSmad2_Smad4_Ski_n,
       pSmad3_Smad4_Ski_n, pSmad3_Smad4_Ski_n, Smad2_Ski_c, Smad3_Ski_c,
       Smad4_Ski_n
11 SnoN: SnoN_c, SnoN_n, pSmad2_Smad4_SnoN_n, pSmad2_Smad4_SnoN_n,
       pSmad2_SnoN_n, pSmad2_SnoN_n, pSmad2_SnoN_n, pSmad3_Smad4_SnoN_n,
       pSmad3_Smad4_SnoN_n, pSmad3_SnoN_n, pSmad3_SnoN_n, pSmad3_SnoN_n,
       Smad4_SnoN_n
12 Smurf1: Smurf1_c, Smurf1_n, Smad7_Smurf1_c, Smad7_Smurf1_n
13 Smurf2: Smurf2_c, Smurf2_n, Smad7_Smurf2_c, Smad7_Smurf2_n
```

References

[1] CP Adams and VV Brantner. **Estimating the cost of new drug development: is it really 802 million dollars?** *Health Aff (Millwood)*, **25**(2):420–8, mar 2006. 2

[2] I Kola and J Landis. **Can the pharmaceutical industry reduce attrition rates?** *Nat Rev Drug Discov*, **3**(8):711–5, aug 2004. 2, 5

[3] DC Swinney and J Anthony. **How were new medicines discovered?** *Nat Rev Drug Discov*, **10**(7):507–19, jun 2011. 2, 3, 5, 6

[4] F Sams-Dodd. **Target-based drug discovery: is something wrong?** *Drug Discov Today*, **10**(2):139–47, jan 2005. 4, 5, 6

[5] J Drews and S Ryser. **Skip to main page content Home OnlineFirst All Issues Subscribe RSS rss Email Alerts Search this journal Advanced Journal Search "¿ Innovation Deficit in the Pharmaceutical Industry**. *Drug Inform Journal*, **30**(1):97–108, jan 1996. 5

[6] SM Paul, DS Mytelka, CT Dunwiddie, CC Personger, BH Munos, SR Lindborg, and AL Schacht. **How to improve R&D productivity: the pharmaceutical industry's grand challenge.** *Nat Rev Drug Discov*, **9**(3):203–14, mar 2010. 5, 6

[7] B Munos. **Lessons from 60 years of pharmaceutical innovation**. *Nat Rev Drug Discov*, **8**(12):959–68, dec 2009. 5

[8] T Ideker, T Galitski, and L Hood. **A new approach to decoding life: systems biology**. *Annu Rev Genomics Hum Genet*, **2**:343–72, 2001. 7, 8

[9] M Schulz and E Klipp. **Introduction to Systems Biology**. *Systems Biology in Psychiatric Research*, pages 81–95, 2010. 7, 8, 9, 11, 20, 21, 22, 25, 26, 27

[10] MA Harris, J Clark, A Ireland, J Lomax, A Ashburner, R Foulger, K Eilbeck, S Lewis, B Marshall, C Mungall, et al. **The Gene Ontology (GO) database and informatics resource**. *Nucleic Acids Research*, **32**(Database issue):D258, jan 2004. 10

[11] M Kanehisa, M Araki, S Goto, M Hattori, M Hirakawa, M Itoh, T Katayama, S Kawashima, S Okuda, T Tokimatsu, and Y Yamanishi. **KEGG for linking genomes to life and the environment**. *Nucleic Acids Research*, **36**(Database issue):D480–4, dec 2007. 10

[12] A Chang, M Scheer, A Grote, I Schomburg, and D Schomburg. **BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009**. *Nucleic Acids Research*, **37**(Database issue):D588–92, nov 2008. 10

[13] I Vastrik, P D'Eustachio, E Schmidt, G Gopinath, D Croft, B de Bono, M Gillespie, B Jassal, S Lewis, L Matthews, G Wu, E Birney, and L Stein. **Reactome: a knowledge base of biologic pathways and processes**. *Genome Biol*, **8**(3):R39, 2007. 10

[14] N Le Novère, B Bornstein, A Broicher, M Courtot, M Donizelli, H Dharuri, L Li, H Sauro, M Schilstra, B Shapiro, JL Snoep, and M Hucka. **BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems**. *Nucleic Acids Research*, **34**(Database issue):D689–91, jan 2006. 10, 44

[15] BG Olivier and JL Snoep. **Web-based kinetic modelling using JWS Online**. *Bioinformatics*, **20**(13):2143–4, sep 2004. 10

[16] F Krause, J Uhlendorf, T Lubitz, M Schulz, E Klipp, and W Liebermeister. **Annotation and merging of SBML models with semanticSBML**. *Bioinformatics*, **26**(3):421–422, doi:10.1093/bioinformatics/btp642., 2010. 10

[17] M Hucka, A Finney, S Hoops, S Keating, and N Le Novère. **Systems biology markup language (SBML) Level 2: structures and facilities for model definitions**. *Nature Precedings*, **58**(2), 2007. 10, 28

[18] A Funahashi, Y Matsuoka, A Jouraku, M Morohashi, N Kikuchi, and H Kitano. **CellDesigner 3.5: a versatile modeling tool for biochemical networks**. *Proceedings of the IEEE*, **96**(8):1254–65, 2008. 10, 37

[19] S Hoops, S Sahle, R Gauges, C Lee, J Pahle, N Simus, M Singhal, L Xu, P Mendes, and U Kummer. **COPASI—a COmplex PAthway SImulator**. *Bioinformatics*, **22**(24):3067–74, dec 2006. 10, 37

[20] EC Butcher, EL Berg, and EJ Kunkel. **Systems biology in drug discovery**. *Nat Biotechnol*, **22**(10):1253–9, oct 2004. 11, 12

[21] CL Stokes et al. **A computer model of chronic asthma with application to clinical studies: example of treatment of exercise-induced asthma**. *J. Allergy. Clin. Immunol.*, **107**:933, 2001. 12

[22] E Lee, A Salic, R Krüger, R Heinrich, and MW Kirschner. **The Roles of APC and Axin Derived from Experimental and Theoretical Analysis of the Wnt Pathway**. *PLoS Biol*, **1**(1):e10, oct 2003. 12

[23] I Swameye, Müller TG, J Timmer, O Sandra, and U Klingmüller. **Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling**. *Proc. Natl. Acad. Sci. U.S.A.*, **100**(3):1028–33, feb 2003. 12

[24] O Radulescu, AN Gorban, A Zinovyev, and A Lilienbaum. **Robust simplifications of multiscale biochemical networks**. *BMC Syst Biol*, **2**(86), oct 2008. 12

[25] K Wegner, A Bachmann, JU Schad, P Lucarelli, S Sahle, P Nickel, C Meyer, U Klingmüller, S Dooley, and U Kummer. **Dynamics and feedback loops in the transforming growth factor signaling pathway**. *Biophys Chem*, **162**:22–34, mar 2012. 12, 13, 44, 53, 55

[26] R Samaga, J Saez-Rodriguez, LG Alexopoulos, PK Sorger, and S Klamt. **The Logic of EGFR/ErbB Signaling: Theoretical Properties and Analysis of High-Throughput Data**. *PLoS Comput Biol*, **5**(8):e1000438, aug 2009. 12

REFERENCES

[27] B Schoeberl, EA Pace, JB Fitzgerald, BD Harms, L Xu, L Nie, B Linggi, A Kalra, V Paragas, R Bukhalid, V Grantcharova, N Kohli, KA West, M Leszczyniecka, MJ Feldhaus, AJ Kudla, and UB Nielsen. **Therapeutically targeting ErbB3: a key node in ligand-induced activation of the ErbB receptor-PI3K axis**. *Sci Signal*, **2**(77):ra31, jun 2009. 12, 71

[28] E Klipp, RC Wade, and U Kummer. **Biochemical network-based drug-target prediction**. *Curr Opin Biotechnol*, **21**(4):511–6, aug 2010. 12

[29] CH Heldin and A Moustakas. **Role of Smads in TGFβ signaling**. *Cell Tissue Res*, **347**:21–36, 2012. 12, 14

[30] P ten Dijke and CS Hill. **New insights into TGF-β–Smad signalling**. *Trends Biochem Sci*, **29**(5):265–73, 2004. 13

[31] DC Clarke, MD Betterton, and X Liu. **Systems theory of Smad signalling**. *Syst Biol (Stevenage)*, **153**(6):412–24, November 2006. 13

[32] JM Vilar, R Jansen, and C Sander. **Signal processing in the TGF-beta superfamily ligand-receptor network**. *PLoS Comput Biol*, **2**(1):e3, jan 2006. 13

[33] B Schmierer, AL Tournier, PA Bates, and CS Hill. **Mathematical modeling identifies Smad nucleocytoplasmic shuttling as a dynamic signal-interpreting system**. *Proc. Natl. Acad. Sci. U.S.A.*, **105**(18):6608–13, may 2008. 13

[34] Z Zi, Z Feng, Chapnick DA, M Dahl, D Deng, E Klipp, A Moustakas, and X Liu. **Quantitative analysis of transient and sustained transforming growth factor-β signaling dynamics**. *Mol Syst Biol*, **7**(492), 2011. 13, 44, 45

[35] B Schmierer and CS Hill. **TGFβ–SMAD signal transduction: molecular specificity and functional flexibility**. *Nat Rev Mol Cell Biol.*, **8**(12):970–82, dec 2007. 13

[36] S Itoh and P ten Dijke. **Negative regulation of TGF-β receptor/Smad signal transduction**. *Curr Opin Cell Biol*, **19**(2):176–84, april 2007. 13

[37] A Nakao, M Afrakhte, A Morén, T Nakayama, JL Christian, R Heuchel, S Itoh, M Kawabata, CH Heldin, and P ten Dijke. **Identification of Smad7, a TGFbeta-inducible antagonist of TGF-beta signalling.** *Nature*, **389**(6651):631–5, oct 1997. 13

[38] K Luo. **Ski and SnoN: negative regulators of TGF-beta signaling**. *Curr Opin Genet Dev*, **14**(1):65–70, feb 2004. 13

[39] A Moustakas and CH Heldin. **Non-Smad TGF-β signals**. *J Cell Sci*, **118**:3573–84, august 2005. 14

[40] D Horbelt, G Guo, PN Robinson, and P Knaus. **Quantitative analysis of TGFBR2 mutations in Marfan-syndrome-related disorders suggests a correlation between phenotypic severity and Smad signaling activity**. *J Cell Sci.*, **123**(Pt 24):4340–50, December 2010. 14

[41] V Sharma, AG Antonacopoulou, S Tanaka, AA Panoutsopulos, V Bravou, HP Kalofonos, and V Episkopou. **Enhancement of TGF- signaling responses by the E3 ubiquitin ligase Arkadia provides tumor suppression in colorectal cancer**. *Cancer Res*, **71**(20):6438–49, oct 2011. 15

[42] Y Ono, T Hayashida, A Konagai, H Okazaki, K Miyao, S Kawachi, M Tanabe, M Shinoda, H Jinno, H Hasegawa, M Kitajima, and Y Kitagawa. **Direct inhibition of the transforming growth factor- pathway by protein-bound polysaccharide through inactivation of Smad2 signaling**. *Cancer Sci*, **103**(2):317–24, feb 2012. 15

[43] B Levin. *Genes VII*. Oxford University Press, Oxford, 1999. 19

[44] E Klipp, W Liebermeister, C Wierling, A Kowald, H Lehrach, and R Herwig. *Systems Biology*. Wiley-Blackwell, Weinheim, 2009. 19, 20, 21, 22, 24, 25, 26, 27, 28

[45] L Petzold. **Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations**. *SIAM J. Sci. Stat. Comput.*, **4**(1):136–48, 1983. 21, 39, 46, 56

[46] AC Hindmarsh, PN Brown, KE Grant, SL Lee, R Serban, DE Shumaker, and CS Woodward. **SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers**. *ACM Trans on Math Software*, **31**(3):363–396, sep 2005. 21

[47] P Waage and C. M. Guldberg. **Studies concerning affinity**. *C. M. Forhandlinger: Videnskabs-Selskabet i Christiana*, 1864. 22

[48] DF Shanno. *Conditioning of quasi/Newton methods for function minimization*, **Mathematics of Computation**, pages 647–56. 1970. 27, 32

[49] S Kirckpatrick. **Optimization by simulated annealing: Quantitative studies**. *Journal of Statistical Physics*, **34**(5):975–86, 1984. 27, 32

[50] M Schulz, Bakker BM, and E Klipp. **TIde: a software for the systematic scanning of drug targets in kinetic network models**. *BMC Bioinformatics*, **10**(344), oct 2009. iii, v, 28, 68

[51] M Schulz. *Identification of Potential Drug Targets in Kinetic Networks Described by Ordinary Differential Equations*. PhD thesis, Freie Universität Berlin, Unpublished. 28, 29, 30, 31, 32, 68, 70

[52] Le Novère. **Model storage, exchange and integration**. *BMC Neuroscience*, **7**(Suppl 1), 2006. 29

[53] JA Nelder and R Mead. **A Simplex Method for Function Minimization**. *The Computer Journal*, **7**(4):308–313, 1965. 32

[54] DE Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989. 32

[55] M Reich, T Liefeld, J Gould, J Lerner, P Tamayo, and JP Mesirov. **Gene Pattern 2.0**. *Nature Genetics*, **38**(5):500–501, 2006. 49, 52

[56] X Duan, YY Liang, Feng XH, and X Lin. **Protein serine/threonine phosphatase PPM1A dephosphorylates Smad1 in the bone morphogenetic protein signaling pathway**. *J Biol Chem*, **281**(48):36526–32, dec 2006. 64

[57] SK Halder, RD Beauchamp, and PK Datta. **A Specific Inhibitor of TGF-β Receptor Kinase, SB-431542, as a Potent Antitumor Agent for Human Cancers**. *Neoplasia*, **7**(5):509–21, may 2005. 65

[58] S Singh, BK Malik, and DK Sharma. **point analysis of metabolic pathways in e. histolytica: A computational approach for drug target identification**. *Bioinformation*, **2**(2):68, 2007. 68

[59] JS Edwards, RU Ibarra, and BO Palsson. **In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data**. *Nat Biotechnol*, **19**(2):125–30, 2001. 68

[60] DS Wishart, C Knox, AC Guo, D Cheng, S Shrivastava, D Tzur, B Gautam, and M Hassanali. **DrugBank: a knowledgebase for drugs, drug actions and drug targets**. *Nucleic Acids Research*, **36**(Database issue):D901, 2008. 69

[61] M Kuhn, D Szklarczyk, A Franceschini, M Campillos, C von Mering, LJ Jensen, A Beyer, and P Bork. **STITCH 2: an interaction network database for small molecules and proteins**. *Nucleic Acids Research*, **38**(Database issue):D552, 2010. 69

[62] K Yang, W Ma, H Liang, Q Ouyang, C Tang, and L Lai. **Dynamic simulations on the arachidonic acid metabolic network**. *PLoS Computational Biology*, **3**(3):e55, 2007. 70