1 Introduction

Using mathematical modeling to address large scale problems in the world of biological regulatory networks has become increasingly necessary given the sheer quantity of data made available by improved technology. In the most general sense, modeling approaches can be thought of as being either quantitative or qualitative. Quantitative methods such as ordinary differential equations or the chemical master equation are widespread in the literature; when the model is well developed, the detail therein can be incredibly informative. However, these methods are not well suited for all applications. Quantitative models require an in depth knowledge of the reaction kinetics and generally fail as the problem size grows. The alternative approach, qualitative models, does not possess the same amount of detail but captures the essential dynamics of the system. In addition, qualitative models have a variety of analysis tools which can be applied regardless of the problem size. Gene regulation, as a sub genre of biological regulatory networks, is characterized by large numbers of interconnected species whose influences depend on passing some threshold, thus, largely sigmoidal behaviors. The application of qualitative methods to these systems can be highly advantageous to the modeler.

In this work, we begin by considering the qualitative framework of Process Hitting, defined briefly in Section 2.1. A highly flexible model, Process Hitting captures the most important dynamics of the system with a relatively simple syntax. The very structure of this syntax lends it to powerful static analysis tools which can be used to answer some of the most important questions about the model such as steady states or reachability without constructing the state space. Realistic models in gene regulation are immense and highly interconnected: even when considering a boolean space, the very enumeration of the possible states of the resulting system creates a combinatorial explosion. This is a frequent obstacle in the field of computer science and has been dubbed the "curse of dimensionality". However, there are some questions for which one must access the underlying probability distribution associated with the Markov transitions of the qualitative model. In addition, gaining access to the probability distribution allows for a qualitative and intuitive analysis of the system as a whole. The most pervasive methods have historically been simulation-based, although there are some instances in which this becomes computationally infeasible. Here, we propose a method to solve the system by treating the Markov equations of a Process Hitting model with numerical techniques. A reduced-basis method, Proper Generalized Decomposition (PGD) can be used to overcome the curse of dimensionality and provide fast, computationally inexpensive solutions to an otherwise intractable problem, as discussed in Section 2.3. In addition, PGD has certain qualities particularly favorable for applications to gene regulatory networks. Unknown parameters can easily be incorporated into the model at the cost of another dimension, as demonstrated in Section 3.2.

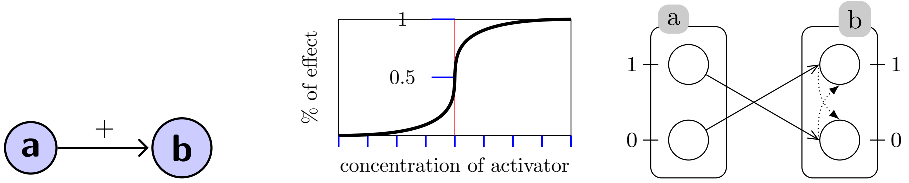

Figure 1: Creating a Process Hitting action. In gene regulation, we consider two kinds of interactions between species: activation and inhibition. If a is an activator of b, it is common to represent this by a signed, directed graph (left). These interactions have a characteristic form: unlike kinetic reactions, activation and inhibition usually depend on the regulator passing some threshold concentration in order to become effective (middle). Process Hitting (right) represents these reactions via actions: a activates b becomes a_1 hits (solid arrow) b_0 to bounce (bashed arc) to b_1. Generalized dynamics attempts to create the most permissive dynamics possible for the directed graph. Therefore, the absence of a effectively acts as an inhibitor, adding the action a_0 hits b_1 to bounce to b_0. Every action can be associated with temporal and stochastic parameters– the reaction rate, for example [**?**].

2 Methodologies

2.1 The Qualitative Model: Process Hitting

Process Hitting is a powerful yet simple tool for the analysis of large regulatory networks. Historically related to the discrete models of Stuart Kauffman [**?**] and René Thomas [**?**], Process Hitting attempts to address problems of scalability in classical modeling methods while maintaining the highest degree of expressiveness possible. Formally a subclass of asynchronous automata, it relies on large degrees of abstraction to describe the system as a whole. All interacting species—whether they be enzymes, genes or proteins —are abstracted as *sorts*. These sorts are then subdivided into *processes*, which could represent concentration levels, spatial configuration, or any other form which has a distinct qualitative impact on the system. Processes interact with one another via *actions*, in which processes *hit* one another to create a *bounce* to some new level of the same sort at a given rate. For gene regulatory networks, processes are often abstractions of relevant concentration ranges, discretized domains of real numbers, and actions represent activation and inhibition reactions. Figure 1 illustrates how to define sorts, processes and actions from a biological understanding of an interaction. Process Hitting relies on the initial construction of the most permissive dynamics, otherwise called *generalized dynamics*, in which no restrictions are placed on the potential behaviors. An example of this can be seen in Figure 1. The general dynamics may then be successively enriched by the addition of *cooperative sorts* in order to best capture some known biological behaviors or eliminate undesirable behaviors. Cooperative sorts represent not species but, rather, the combined effects when multiple regulators interact cooperatively on a single target. These sorts are the combined space of the original species, thus must be updated such that the current state of the cooperative sort is compatible with the current state of each of its components. A visual explanation of the construction of a cooperative sort and its refinement of a Process Hitting model can be found in Figure2.

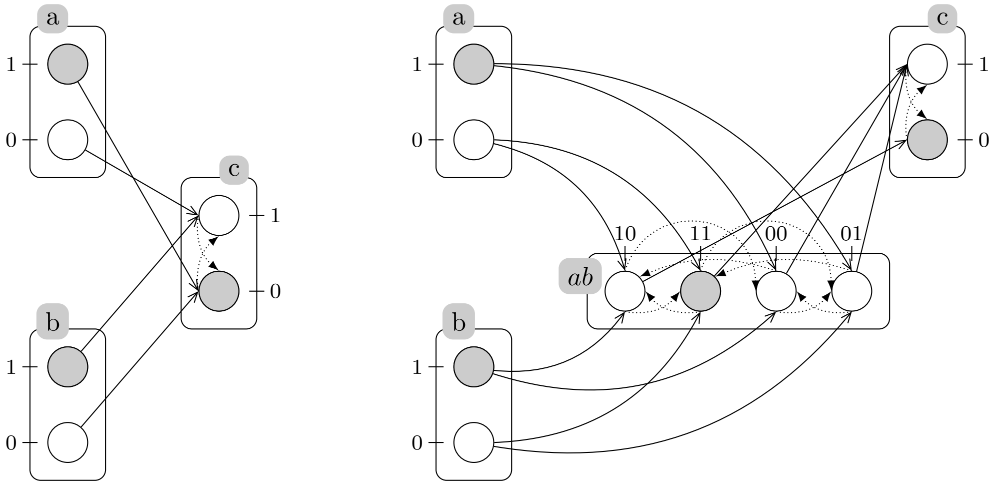

Figure 2: Refinement of a model via Cooperative Sorts. Here, a is an activator of c while b inhibits c. The generalized dynamics of the system have been constructed on the left. But what should happen in the case that both a and b are present? According to the left hand model, the system will oscillate. If we know more about how the system should function, however, we would like to be able to include this information into our model. With general dynamics, we are unable to express logical gates in which multiple species exhibit deterministic combined effects on a target, such as $a \wedge \neg b$, or the presence of the activator without the presence of the inhibitor. In order to add this combined interaction and eliminate the oscillatory behavior, we must refine the Process Hitting model with a cooperative sort, ab. This sort will handle the interactions of a and b on c while leaving the original species to interact with other elements as before. In exchange, more actions must be added such that a and b can effectively update ab so that it truly reflects the current state of both elements. In our example, $ab_{1,1}$ will not interact with c_0, thus c remains inactive.

Although this is a very simplistic representation of the inner kinetics of a biological process, Process Hitting semantics allow us to easily model interactions with only partial knowledge of the logical functions encoded therein and pave the way for powerful static analysis techniques in order to study fixed points, reachability and cut sets which determine minimum criteria for reachability, in spite of the present combinatorial explosion [**?**, **?**]. Examples of Process Hitting at work can be found in Section 3, where we use static analysis to compare the fitness of the generalized dynamics model with that of the refined model. Furthermore, these tools are freely available online in a software called PINT **??**. We will not attempt to expound completely upon the details of Process Hitting here but, rather, point those interested towards [**?**] for a formal and thorough introduction to the modeling framework. As we progress to a biological application in Section 3, greater clarity will be given to the concepts described above, including the relevance of cooperative sorts and the power of static analysis.

2.2 Treating Qualitative Systems with Numerical Techniques

In order to address Process Hitting's global results, that is, the full and complete description of the systems behavior given an initial condition, we must consider the framework in a stochastic context. Process Hitting actions move the system from one state, z, to another,$\hat{z}$, at a given propensity which depends on only the current state and time, or $a_j(z,t)$. As a memoryless random walk, each action corresponds to a Markov equation which tracks the net change in the probability of existing at a certain state and time:

$$\frac{\partial \Phi(z,t)}{\partial t} = \sum_j a_j(\hat{z},t)\Phi(\hat{z},t) - \sum_k a_k(z,t)\Phi(z,t)$$

The result is a system of linear, time dependent, partial differential equations, defined given an initial condition. The solution is a multivariate Bernoulli distribution in which exactly one of the K outcomes is successful, or 1-in-K. This differs substantially from the multivariate Gaussian distribution in that all of the indecies are permutable, a particularly challenging characteristic of the Process Hitting structure in terms of numerical solvers. Some of the most famous and broadly used techniques for addressing problems such as these have been simulation based. While this does avoid constructing the full state space, simulation can become computationally expensive with respect to run-time and available memory. An alternative approach is the direct application of a numerical method to the Markov equations. Here, we propose Proper Generalized Decomposition (PGD) as an effective and well suited technique for gene regulatory networks.

2.3 Proper Generalized Decomposition

Proper Generalized Decomposition [**?**, **?**] is a multi-linear numerical solver which assumes that the target, in this case, the probability distribution, can be written as a sum of a product of separable functions of the interacting species, $F^j(z_i)$ $i = \{1 \cdots N_{sp}\}$, and time, $F^j(t)$:

$$\Phi(z,t) \cong \sum_{j=1}^{M} F_1^j(z_1) \cdots F_2^j(z_2) \cdots F_{N_{sp}}^j(z_{N_{sp}}) \cdot F_t^j(t)$$

PGD is performed iteratively, starting at some arbitrary guess and searching for sets of functions, one vector at a time, which will minimize the residual of the running sum. These functions are colloquially called "modes", however, since the only objective is the reduction of the residual, there is no underlying notion that they represent the greatest source of variance, as is the case with Principal Component Analysis. Although the accuracy increases with every addition, we assume that only a limited number, M, of sets of functions are needed to capture the behavior of the system. If we consider a network of N_{sp} species with N possible levels, the resulting dimensionality is the M sum of N_{sp} functions of size N, or $M(N \times N_{sp})$ in contrast to the original N_{sp}^N. We have not changed the state space but, rather, re-ordered it such that only one $N \times 1$ vector, usually on the scale of 2 or 3, must be addressed at any given time, see Figure 3. Since all operations can be performed by canonical techniques and are highly parallelizable, iterations are generally fast and computationally inexpensive.

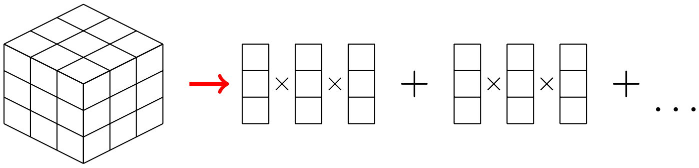

Figure 3: Decomposition of a state space. This illustration shows how a multidimensional space, for example, a cubic space of three dimensions, 3^3 can be decomposed into the product of the individual dimensions, 3×3. This mathematical property is exploited by PGD in that we search for the individual vectors which are of relatively small size, never touching the full state space. In such a way, we move from N_{sp}^N to $M(N \times N_{sp})$, as M sets of these vectors must be found.

3 Application to a Biological Network

It is easier to understand the concepts of Process Hitting and PGD, as well as to see their individual and combined benefits, when seen "in action" in the context of a realistic application towards a gene regulatory network. Here, we investigate a medium scale model of the ErbB signaling pathway which regulates a cells transition from G1 to S life phase, an important checkpoint which determines whether a cell should divide, delay division or enter a quiescent state. Over expression of ErbB is associated with many kinds of cancer, and drugs which target it and its receptor are common treatments for breast, lung and colon cancers. The directed graph for this network was taken from [**?**], where twenty species interact according to Boolean rules. The directed graph can be found in the appendix for reference. We begin our application by constructing a Process Hitting model from this Boolean predecessor, taking the most permissive, generalized dynamics, followed by its refinement via the incorporation of cooperative sorts, the process for which was seen in Figure 2. The impact that this refinement has on both the static analysis and application of PGD will be investigated, both in terms of expressiveness and complexity. Finally, the potential of PGD's capacity to easily incorporate model parameters as extra coordinates will be demonstrated by taking many potential values for the rates of two reactions in the directed graph.

The translation of a Boolean model to the generalized dynamics of Process Hitting is relatively straight forward, as shown in Figure 1: the absence of an activator effectively serves as an inhibitor and vice versa. The formal relationship between Boolean networks and Process Hitting can be found in [?]. At this point, we would like to investigate the model to see if it adequately reflects our biological understanding of the system as a whole: are experimentally demonstrated states reachable, are impossible states unreachable, and are there fixed points if steady state behaviors exist? These questions constitute sanity checks, making sure our model is not essentially flawed from the beginning. The structure of the system, see appendix, suggests two species of experimental interest: EGF as an input, having no predecessor, and pRB as output, having no successor. Using these two species, we can easily formulate simple reachability criteria in order to perform sanity checks on our model. We consider a system "at rest", in which all components begin in their inactive state. If no changes are made on the input protein, EGF when it is inactive, we expect that the system will remain at rest and that no change is to occur in the output protein. However if EFG is introduced, the signal should be able to propagate to the output, pRB. In order to be a feasible model, the system must pass these two criteria. Results from static analysis, shown in Table 1 provide good evidence that the generalized dynamics are too permissive and do not accurately capture the biological behaviors which are essential for a functioning model: not only are we are unable to find any fixed points within the system, of which we do expect to find at least one, but the protein pRB may become sporadicly activated in a globally inactive system, failing the first sanity check. Therefore, we must refine the model, incorporating the suggested logical gate rules from [?] via cooperative sorts. In doing so, we recapture these vital phenomena, finding three fixed points and passing both sanity checks. These results were obtained in a matter of seconds, using simple commands in freely available software, allowing us to efficiently alter our model before investing time in more computationally expensive analysis.

3.1 Cooperative Sorts in the context of PGD

The Markov Equations of the Process Hitting actions provide a system of PDEs to which we can apply PGD. Each species occupies a dimension of the state space. With two processes to each sort, the final problem is of size 2^{20}, or over one million possible states. The underlying probability distribution is a function of these species and time. Our goal is to approximate this solution by

Model	**Fixed points**	EGF absent	EGF present
Gen. Dynam.	0	Fail	Pass
Refined	3	Pass	Pass

Table 1: Results for ErbB models using generalized dynamics and a refinement with cooperative sorts. Here, the two models were tested using two three sanity checks related to our biological understanding of the system: the presence of fixed points, the lack of impossible behaviors and the presence of demonstrated behaviors. In order to be considered a functioning model, pRB should remain at rest when the system is universally inactive, including the absence of input protein EGF. However, in the presence of EGF, a signal should be able to propagate through the system, potentially activating pRB. We see that, while the generalized dynamics were able to propagate a signal from EGF to pRB (EGF present), it was not able to prevent sporadic activation of pRB in a system at rest (EGF absent), nor find any fixed points.

a summation of separable functions

$$\Phi(z,t) \cong \sum_{j=1}^{M} F_1^j(EGF) \cdots F_{N_sp}^j(pRB) F^j(t)$$

In the case of a Process Hitting containing only the generalized dynamics, this is an appropriate and accurate method. However, once cooperative sorts are incorporated into the qualitative model, the cooperating species can no longer be represented by separable functions. To satisfy the enriched model, we may simply combine those dimensions which participate in cooperative sorts. While this does create vectors which grow exponentially with each added species, it is biologically implausible that more than three or four species would participate in a cooperative influence on a single target. Therefore, we can expect this growth to be cut short long before the dimension of a cooperative sort becomes too large. As we combine the state spaces so that they reflect their cooperative sorts, the error associated with PGD solutions as compared to the solution obtained from simulation techniques decreases, figure 4. But what is to be done when one species participates in multiple cooperative sorts? The species cannot be represented twice in the decomposition, so we cannot construct two separate entities for the cooperative sorts as we would like. Rather, if we simply combine the two cooperative sorts into a single element, we return to the most accurate representation of the system, as each species is only represented once, but any non-separable behavior can be taken into account. Again, while it is possible to experience exponential growth in the combination of cooperative sorts, it is very rare biologically that one species would be participant in more than a handful of interactions.

The solutions that we obtain from PGD are approximations of the full probability distribution corresponding to the Markov Equations created by Process

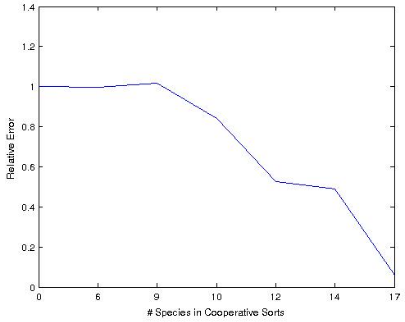

Figure 4: Error inherent in the model decreases with the increased incorporation of combined state spaces. In this case, error is judged as the Euclidian distance from the averaged results of simulation for 1000 trials.

Hitting. From these probability distributions, we are able to make fast analysis of global behaviors of the system: rather than being limited to asking the questions answerable using static analysis, a modeler can watch the system evolve through time and make general statements on the qualitative behavior.

3.2 Incorporation of Unknown Parameters

It is often the case, especially in a growing field such as geneomics, that elements of a regulatory network are disputed or unknown. Researchers may come to very different conclusions about the parameters which fit a particular system. With simulation techniques, each new set of parameters requires a full repetition of all of the trials, limiting the modeler and leading to ad hoc choices made for the sake of feasibility. However, PGD offers a simple way of incorporating these unknown parameters directly into the model, making it possible to obtain an approximative solution for a range of values all at once[**?**]. The parameter is encoded as one of the separable spaces and is included at the cost of one dimension added to the overall solution space. For our example, perhaps one of the regulating reactions is difficult to study separately from the system as a whole, say, interactions involving p27 and p21. Unlike the first half of the directed graph which is simply an activation cascade, these proteins are involved in both inhibiting and activating relationships, so changes to their rate laws should more greatly influence the final expression of pRB. We would like to incorporate many potential values of the action firing rate r into our model, anywhere between two times faster and two times slower than the other reactions in the system. In order to do so, our decomposition of $\Phi(z,t)$ is changed slightly in order to accommodate the parameter for the range of possible values discretized into tenths:

$$\Phi(z,t,r) \cong \sum_{j=1}^{M} F_1^j(z_1)\cdots F_{N_{sp}}^j(z_{N_{sp}})F_t^j(t)F_r^j(r)$$

While simulation run time grows linearly with each element, 40 times longer since there are 40 values in the discretization, to obtain a result, we are able to derive a solution in relatively equal time using PGD. In figure 5, we see three solutions for the protein pRB given different values of the parameter r.

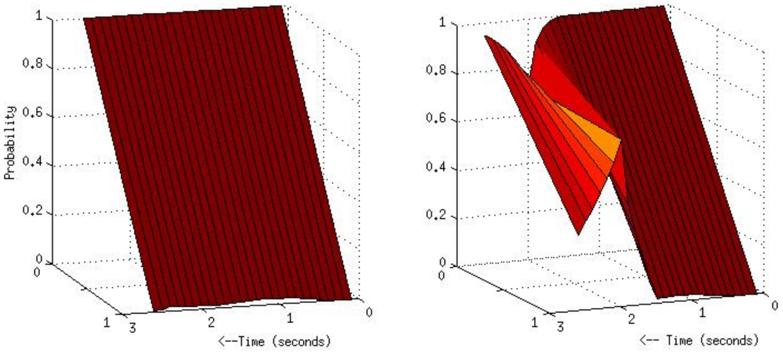

Figure 5: Change in the behavior of pRB given different values for the firing rates of any interaction involving the proteins p21 or p27. We imagine a scenario in which these behaviors cannot be studied separately from the system at large, and therefore must be estimated or incorporated into the model as parameters themselves. Considering that the value could be anywhere between two times slower or two times faster than the rates of the other interactions, we include functions of r, $F(r)$, into the decomposition of Φ. Here, we see two such samples of the resulting PGD solution for the output protein pRB: 1.5 times slower (left) and 1.5 times faster (right). In the right hand case, we can see that the dynamics allow the protein to become active in a shorter amount of time, a non-obvious result as inhibiting interactions of p27 and p21 also become faster.

4 Conclusion and Final Remarks

In the case of gene regulatory networks, there are many reasons why a modeler might choose the application of qualitative methods, one of which is Process Hitting. Process Hitting offers many advantages for large scale, which are often the more realistic, systems in the form of static analysis tools. These analysis tools alone, however, cannot provide the complete and intuitive solution of the system as a full probability distribution for each state over time. By translating Process Hitting actions to Markov equations, we are able to treat a system of PDEs directly. Proper Generalized Decomposition has proven efficient in solving Process Hitting models. As opposed to simulation techniques, which have been historically been the preferred methodology, PGD can provide full solutions, including multiple unknown paramters, with a single run. Here, we have shown some of the potential of this method, applying a combination of static analysis and numerical tools in order to maximize the expressiveness and understnading of a qualitative model. Only the basic elements of Process Hitting have been incorporated into the Markov equations considered, that is, actions with simple rate laws. Including temporal and varied stochastic features into these equations would further increase its potential.

A The ErbB Signaling Pathway

For this work, we used a Boolean model of the ErbB signaling pathway for the regulation of G1/S cell cycle transition as developed by [**?**]. In this article, the authors began by constructing a model from the literature, then proceeded to refine the model via network reconstruction. Although these refinements proved useful in the selection of novel targets for gene therapy, we would like to focus on the initial derivation of the model in which all reactions correspond to cited regulations. However, we will use the logical rules suggested within this article for the refinement of the Process Hitting model via cooperative sorts, shown in Table 2.

EGF (epidermal growth factor) binds to ErbB receptors, of which there are four structural variants, three thought to be involed in this network. These receptors are functional when they form heterodimers, excluding ErbB1 which is able to function as a homodimer as well. Functional receptors transmit signals to AKT1, an apoptosis-inhibiting transmitter, and MEK1, a protein kinase. Along with transcription factors c-MYC and ER-α, these entities downregulate kinase inhibitors p21 and p27 while upregulating the cyclins (CycE1 and CycD1) needed to activate their respective cyclin dependent kinases (CDK). These CDKs will work to phosphoralize, and therefore inactivate, the retinoblastoma protein (pRB). Only when this protein is inactive can the E2F group of transcriptional factors required for DNA replication and, therefore, cell proliferation. Although the interaction between CDKs and pRBs is inhibitive, we have kept the activations as indicated by the authors, using pRB as a proxy for it's following and more interesting product, E2F. In addition, we have included the logical rule proposed for Cyclin D presented in their work.

Target	**Logical Rule**
ERBB1-2	ERBB1 $\wedge ERBB2$
ERBB1-3	ERBB1 $\wedge ERBB3$
ERBB2-3	ERBB2 $\wedge ERBB3$
IGF1R	(ER-α $\vee AKT1) \vee \neg ErbB2-3$
ER-α	AKT1 $\vee MEK1$
c-MYC	AKT1 $\vee MEK1 \vee ER{-}\alpha$
AKT1	ErbB1 $\vee ErbB1-2 \vee ErbB1-3 \vee ErbB2-3 \vee IGF1R$
MEK1	ErbB1 $\vee ErbB1-2 \vee ErbB1-3 \vee ErbB2-3 \vee IGF1R$
CDK2	CycE1 $\wedge \neg p21 \wedge \neg p27$
CDK2	CycD1 $\wedge \neg p21 \wedge \neg p27$
CycD1	ER-α $\wedge c-MYC \wedge (AKT1 \vee MEK1)$
p21	Er-α $\wedge \neg AKT1 \wedge \neg c-MYC \wedge \neg CDK4$
p27	Er-α $\wedge \neg CDK4 \wedge \neg CDK2 \wedge \neg AKT1 \wedge \neg c-MYC$
pRB	(CDK4 $\wedge CDK6) \vee (CDK4 \wedge CDK6 \wedge CDK2)$

Table 2: The proposed logical rules for species with more than one regulator.

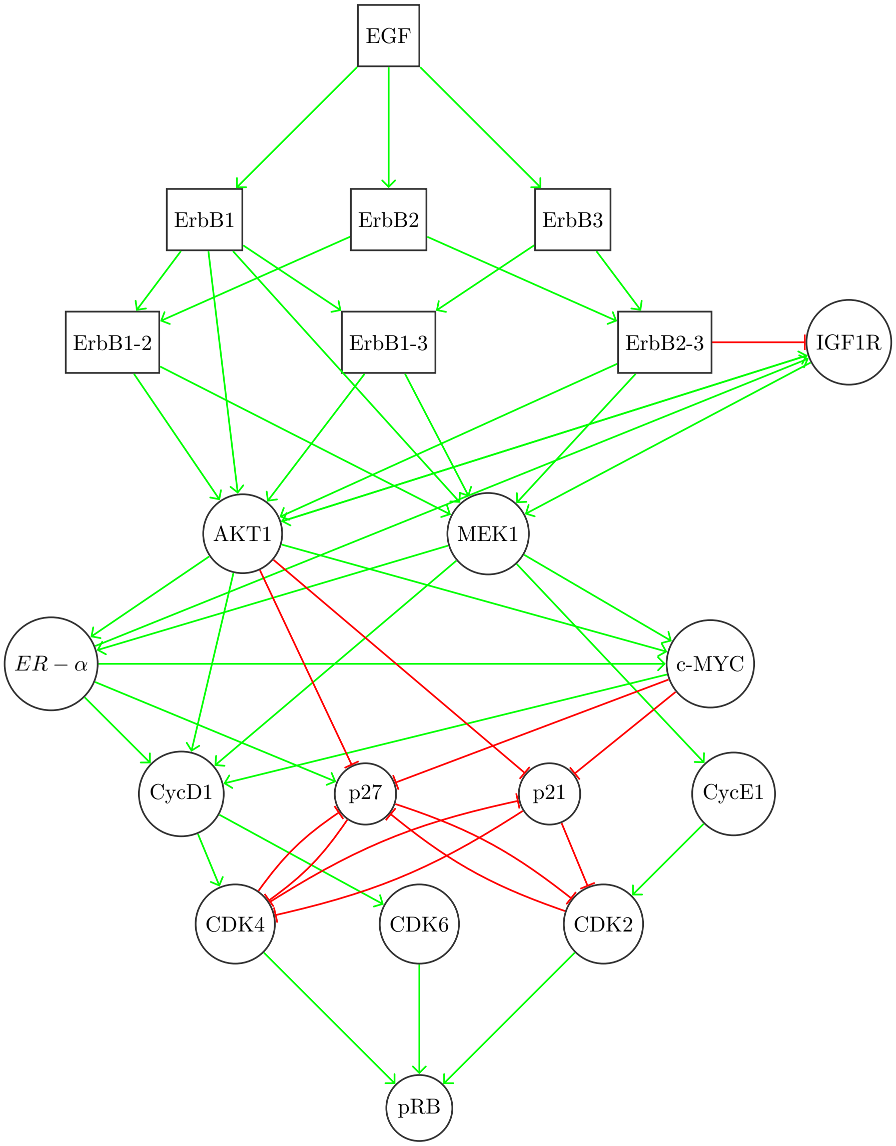

Figure 6: The interaction graph for ErbB mediated G1/S cell cycle transition. Here, elements directly related to the ErbB signaling portion of the network are represented by boxes, while the elements related to kinase activity are represented by circles. Activation interactions are shown in green arrows and inhibition in red blunted arrows. Since this is the initial, most basic network derived from the literature, no combined effects requiring Boolean logic gates are shown.