Generating Synthetic Signaling Networks for in Silico Modeling Studies

Jin Xu[a], H. Steven Wiley[b], Herbert M. Sauro[a]

[a]*Department of Bioengineering, University of Washington, Seattle, WA, USA*
[b]*Environmental Molecular Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA*

Abstract

Predictive models of signaling pathways have proven to be difficult to develop. Reasons include the uncertainty in the number of species, the complexity in species' interactions, and the sparseness and uncertainty in experimental data. Traditional approaches to developing mechanistic models rely on collecting experimental data and fitting a single model to that data. This approach works for simple systems but has proven unreliable for complex systems such as biological signaling networks. For example, uncertainty and sparseness of the data often result in overfitted models that have little predictive value beyond recapitulating the experimental data itself. Thus, there is a need to develop new approaches to create predictive mechanistic models of complex systems. However, to determine the effectiveness of any new algorithm, a baseline model is needed to test its performance. To meet this need, we developed a method for generating artificial synthetic networks that are reasonably realistic and thus can be treated as ground truth models. These synthetic models can then be used to generate synthetic data for developing and testing algorithms designed to recover the underlying network topology and associated parameters. Here, we describe a simple approach for generating synthetic signaling networks that can be used for this purpose.

1. Introduction

In this article we describe a Julia [1] package that can be used to generate synthetic signaling networks. Such networks can be used as a basis to test novel algorithms designed to identify the topology and parameters of real signaling networks from biological data. This is particularly important for

building predictive models of signaling networks involved in diseases, such as cancer [2, 3]. Methods for developing predictive models of biochemical pathways have not changed significantly over the last 60 years. Many of the commonly used techniques have been translated directly from other disciplines where systems tend to be much simpler. One approach is to collect experimental data and fit it to a single model using a suitable optimization algorithm [4, 5]. Given the large number of state variables and parameters present in signaling network models, the availability of limited data results in severe overfitting and non-identifiability of parameters [6, 7], even assuming that the topology of the model is correct. Such models fail to generalize and often have poor predictive value. This problem is by no means restricted to just biological systems but applies to any complex system, such as weather forecasting [8], climate models [9, 10], financial models [11] or hydrodynamic models [12]. Given the difficulties and importance in being able to develop predictive models of such systems, there is a general need for the development of novel approaches that take into account the uncertainties in our knowledge and ability to make measurements that will generate models with the most predictive power. However, evaluating the effectiveness of any new model generating or parameterization algorithm requires the availability of "ground truth" models against which algorithm output can be compared. One approach is to generate artificial synthetic networks that are reasonably realistic that can serve as ground truth models. Such models can be used to generate artificial "experimental" data that can be used to test an algorithm's ability to recover the original model and species parameters. In this article, we describe a computational approach for generating synthetic signaling networks.

Biological signaling pathways [13] are information-processing networks that are used by cells to translate external signals and cues into appropriate cell actions, such as cell growth, differentiation, movement, death or metabolic activity. Signaling pathways are based on the interaction of a network of proteins through a limited number of processes, such as phosphorylation, protein complex formation and targeted cleavage, degradation or synthesis [14]. A cell's response to external signals is typically mediated by specific receptor proteins, which carry the information into cells through a series of complex steps that amplify and process the signal prior to its output to the effector function. The amplification and signal processing steps frequently use enzymatic steps, such as protein phosphorylation or proteolysis. Signaling pathways also frequently display multiple feedback loops that

regulate the dynamic behavior of the network and its information processing ability. In many cases the role of feedback loops and the overall topology of the signaling pathways is unclear.

The importance of cell signaling to cancer is well known [15, 16] and an understanding of how signaling networks contribute to the disease would clearly be useful. If reliable predictive dynamic models of signaling pathways could be developed, then more rational drug design and targeting would be possible. However, developing predictive mechanistic models of signaling pathways is difficult due to the large number of interactive components (including potentially unknown interactions) and the sparsity of suitable data to calibrate them. To address this need, we have been developing perturbation-based approaches to infer the underlying topology of signaling networks [17] . Synthetic networks would be useful in this effort, but they must recapitulate the biophysical constraints that govern signaling pathways. Thus, we set out to define a general approach to generating synthetic signaling networks that appropriately resemble natural ones.

2. Methods

To generate synthetic networks, we first created a list of unit processes that are present in most signaling networks. These are shown in Figure 1 and includes catalyzed transformations (A), three binding and unbinding reactions (B-D), and two phosphorylation/dephosphorylation units (E, F) which include single (E) and doubly-phosphorylated (F) motifs. The three binding and unbinding reactions (B-D) follow mass-action kinetic rate laws. The other three units (A, E, F) follow reversible Michaelis-Menten rate laws. Their corresponding rate equations are shown in Table 1. All rate constants and species concentrations are assigned randomly.

Many proteins in signaling networks are found in complexes with other proteins. As a result, the algorithm starts by defining a finite set of monomeric proteins and uses these in combination with the reaction process shown in Figure 1 to generate a signaling pathway that can consist of multi-protein complexes. Reaction process from Table 1 are selected at random with predefined probabilities. Each signaling pathway also has a designated input and output species and the network is grown between these two points. The user can also specify the number of species and the maximum number of reaction units that should be included in the final network. In this way arbitrarily complex networks can be generated.

Unit type	**Rate laws**
uni-uni (A)	$v_A = \frac{C(k_f \cdot A/K_A - k_r \cdot B/K_B)}{1 + A/K_A + B/K_B}$
uni-bi (B)	$v_B = k_f \cdot A - k_r \cdot B \cdot C$
bi-uni (C)	$v_C = k_f \cdot A \cdot B - k_r \cdot C$
bi-bi (D)	$v_D = k_f \cdot A \cdot B - k_r \cdot C \cdot D$
Single phosphorylation dephosphorylation cycle (E)	$v_{E1} = \frac{C(k_{f1} \cdot A/K_{A1} - k_{r1} \cdot B/K_{B1})}{1 + A/K_{A1} + B/K_{B1}}$ $v_{E2} = \frac{D(k_{f2} \cdot B/K_{B2} - k_{r2} \cdot A/K_{A2})}{1 + B/K_{B2} + A/K_{A2}}$
Dual phosphorylation dephosphorylation cycle (F)	$v_{F1} = \frac{D(k_{f1} \cdot A/K_{A1} - k_{r1} \cdot B/K_{B1})}{1 + A/K_{A1} + B/K_{B1}}$ $v_{F2} = \frac{D(k_{f2} \cdot B/K_{B2} - k_{r2} \cdot C/K_{C2})}{1 + B/K_{B2} + C/K_{C2}}$ $v_{F3} = \frac{E(k_{f3} \cdot C/K_{C3} - k_{r3} \cdot B/K_{B3})}{1 + C/K_{C3} + B/K_{B3}}$ $v_{F4} = \frac{E(k_{f4} \cdot B/K_{B4} - k_{r4} \cdot A/K_{A4})}{1 + B/K_{B4} + A/K_{A4}}$

Table 1: The rate equations of the six types of units composing the artificial random networks. The three binding and dissociation reactions (B-D) follow mass-action kinetic rate laws. The other three units (A, E, F) follow reversible Michaelis-Menten rate laws.

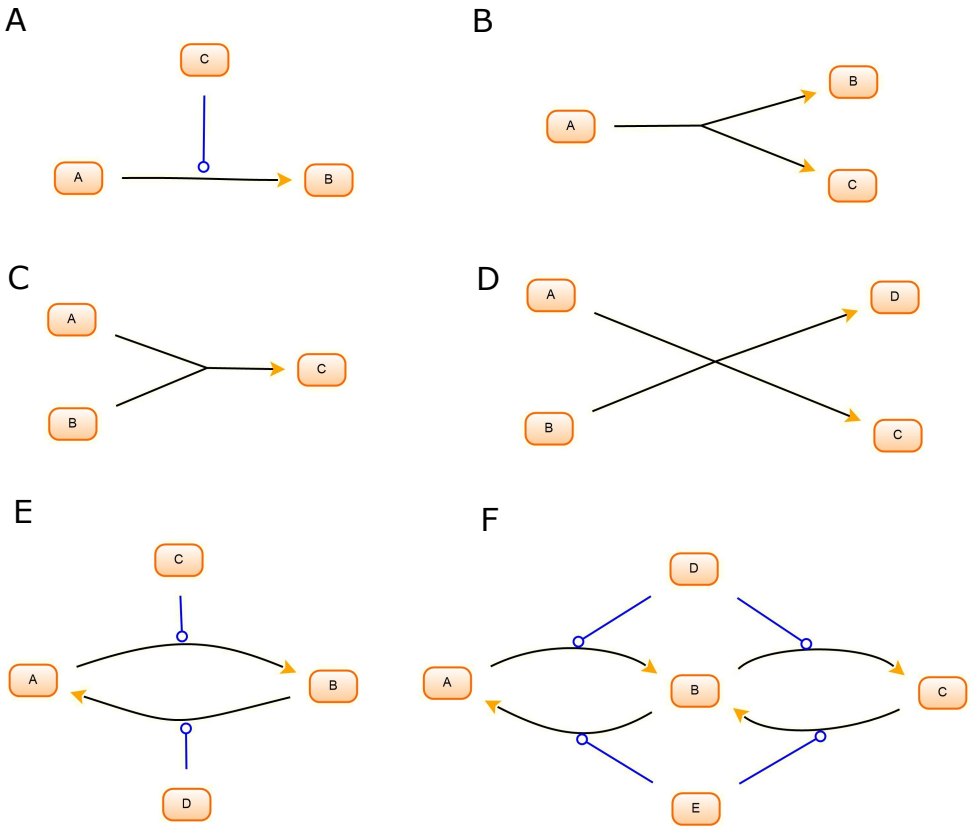

Figure 1: The synthetic random networks are composed of six types of reactions: (A) catalyzed uni-uni, (B, C) binding and unbinding reactions, (D) exchange reactions, (E) catalyzed phosphorylation-dephosphorylation, and (F) dual phosphorylation-dephosphorylation cycles.

However, not all initially generated networks are viable or useful. Thus, the algorithm imposes structural constraints that must be followed. These include requiring that all species must be connected to a path in the network that connects the input to the output species, which ensures that no isolated species exist. In addition, this prevents reaction fragments from being isolated from the main body. There are also dynamic constraints in the network generation system. Networks that cannot reach a steady state are excluded. We have also found that although some networks appear complex and connected, perturbations to the input fail to propagate to the output. Because such models by definition lack the ability to process information,

they were also excluded. We have also excluded networks where the input species is directly connected to the output species because those networks lack the ability to process information.

The code to generate the synthetic networks was written using the Julia Language (`https://julialang.org/`) and the simulations were done using libRoadRunner library [18] so that we could also export SBML. See appendix for details.

3. Results

Figure 2 illustrates an example of a randomly generated signaling network. It includes all six types of the chemical reaction processes in Table 1. Species are labelled `S1` to `S15`. These include one uni-uni process: `S13` $\rightarrow$ `S2` catalyzed by `S11`; one bi-uni reaction: `S14 + S15` $\rightarrow$ `S9`; one uni-bi reaction: `S12` $\rightarrow$ `S8 + S10`; one bi-bi reaction: `S8+S10` $\rightarrow$ `S4 + S13`; one single phosphorylation-dephosphorylation cycle: `S7` $\leftrightharpoons$ `S_out` catalyzed by `S2` and `S4`; and a single dual phosphorylation-dephosphorylation cycle `S3` $\leftrightharpoons$ `S1` $\leftrightharpoons$ `S11` catalyzed by `S6` and `S9`.

Additional random signaling networks are shown in Figure 3. As shown, all the random signaling networks have 15 species in addition to input and output species, with a limit of 15 reactions in total. The time taken to generate a single network that satisfies the constraints described in the Methods section can range from just under a minute to ten minutes for large networks (Table 2). Figure 4 illustrates some simulations where we shown how the concentrations of the output species reach a steady states. Figure 5 shows a time-course simulation of the output species, `S_out`, where there is a perturbation to the input species. All computations reported in Table 2 were done using an Intel i7 9700 processor running at 3.00 GHz with 32GB RAM using Windows 10.

Species #	15	20	25
Time (sec)	99.71 $\pm$ 127.31	247.80 $\pm$ 200.23	550.92 $\pm$ 529.40

Table 2: The time taken to generate and find a qualified random signaling network depends on the size of the network. The size of the random network is represented by the number of species involved (input and output species are not included). The errors represent the standard deviations from ten independent runs.

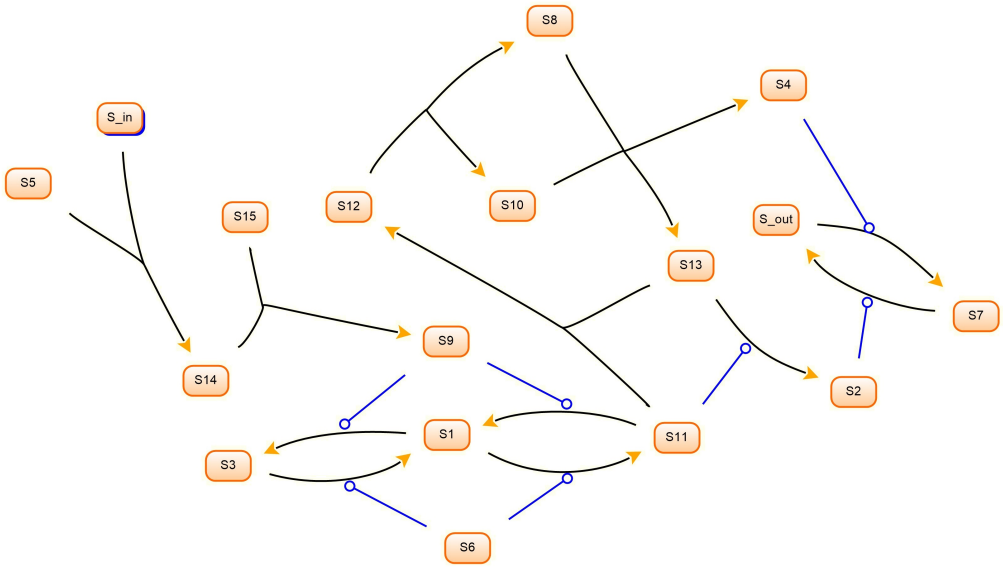

Figure 2: Random signaling networks with 15 species in addition to an input and output species (`S_in` and `S_out`) and seven catalytic interactions are shown in blue lines with a circle at the end.

4. Discussion

We have developed a Julia package that allows users to computationally generate artificial random signaling networks that can be used to test novel algorithms for parameter fitting, topology mapping and other analyses. At present we do not include any kind of sequestration in the phosphorylation cycles. Previous work has shown that sequestration of kinases and phosphatases [19] can have a significant effect on signaling network behavior. Such effects cannot be modelled with the current version of the software. We have not examined whether our generated networks have similar graph metrics to networks found in nature, which is another area that should be investigated. A more thorough investigation into the potential dynamics of our random networks needs to be carried out in order to investigate how the behavior compares to natural networks. We have also not imposed biophysical constraints on species parameters. Because network behavior is a function of both topology and species parameters [20], such constraints could reduce the number of topological configurations that could produce viable networks. Finally, further analysis of which constraints must be either imposed or relaxed

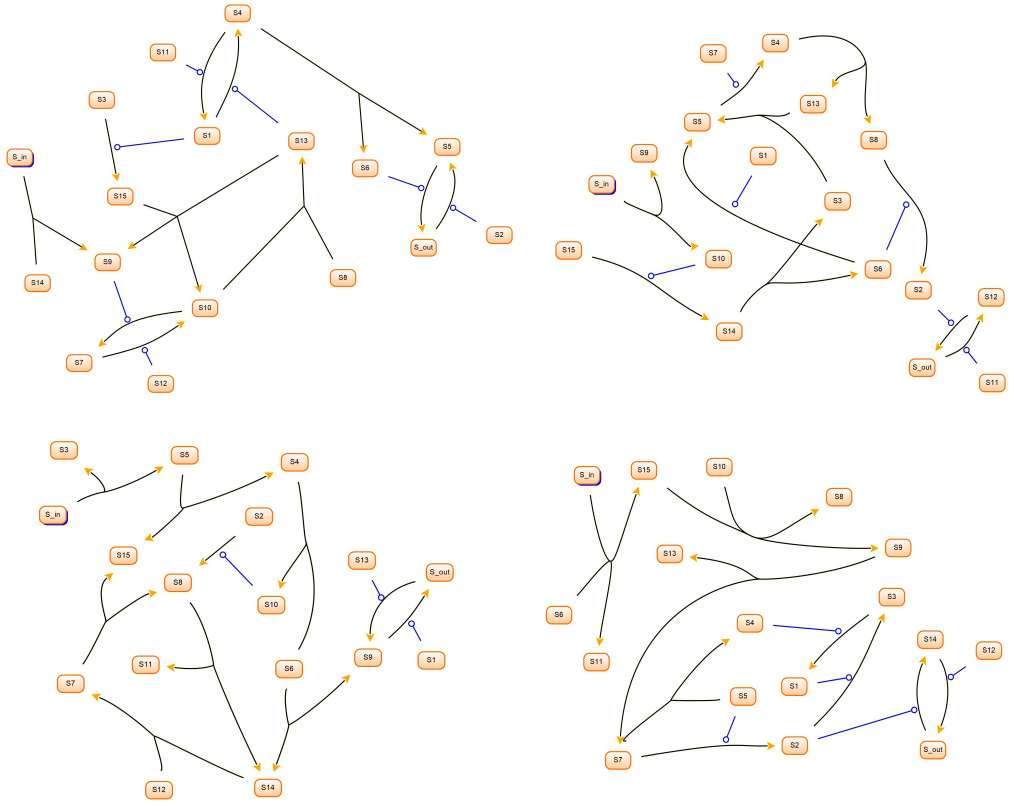

Figure 3: Four additional examples of random signaling networks with 15 species in addition to input and output species, S_in and S_out.

to generate realistic network behavior could reveal important principles that real biological networks follow during their evolution.

5. Author contributions

JX wrote the code, generated the results and wrote the initial article draft. HSW assisted in writing the article. HMS conceived the idea and assisted in writing the article.

6. Acknowledgement

This work was financially supported by the National Institute of General Medical Sciences and National Cancer Institute under grant number

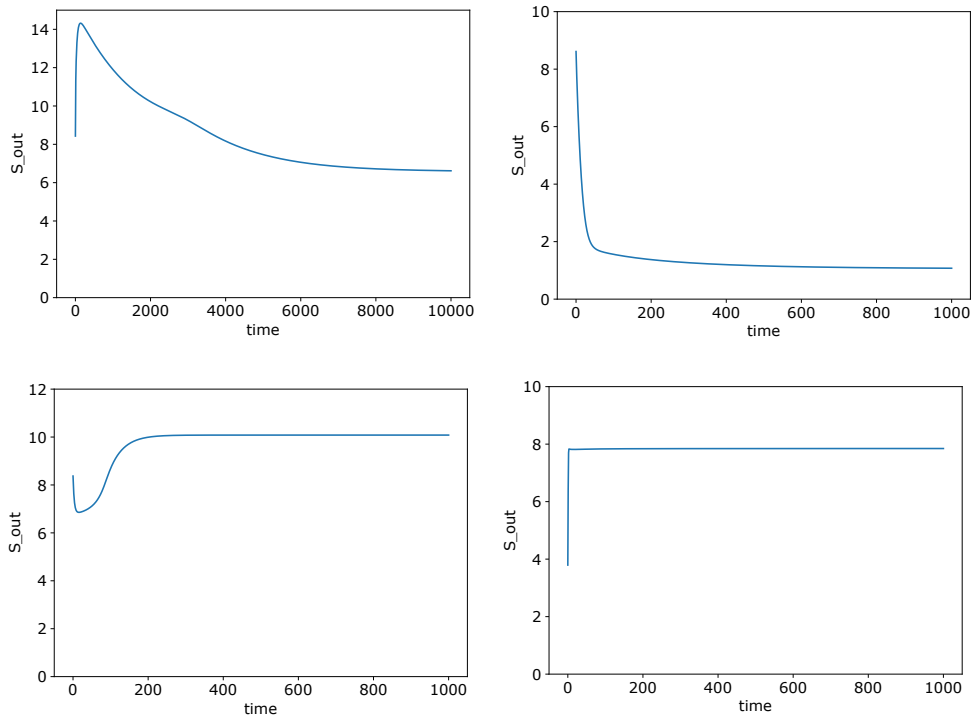

Figure 4: Time-course simulations for four randomly generated networks made up of 15 species in addition to input and output species. The plots show how the concentrations of the output species (S_out) reach their steady states.

GM12303201 and U01CA242992. The authors are responsible for the content, which does not necessarily represent the opinions of the National Institutes of Health. We thank Luke Y. Zhu for assisting with development of some of the early code.

Appendix A. Code availability

The code to generate the artificial signaling random networks is available at `https://github.com/sys-bio/aritificial_random_signaling_network`. This package was implemented in Julia 1.2 on Windows 10. To use the package, first install Julia by going to the website: `https://julialang.org/`,

Once you have the Julia console open, make sure you have `StatsBase` installed by typing `using Pkg` followed by `Pkg.add("StatsBase")`. Note

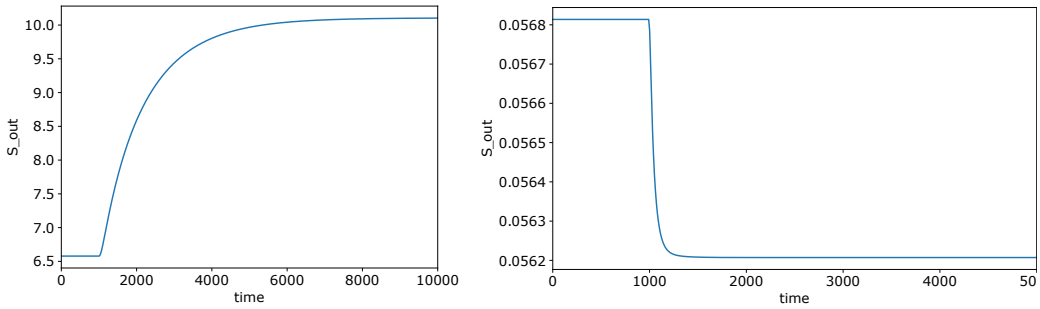

Figure 5: Time-course simulations of concentrations for the output species, `S_out`, as a result of a perturbation to the input species. The two subplots correspond to different artificial random networks with 15 species in addition to input and output species.

there shouldn't be a space between the `add` and the first parenthesis. This is a package required for random number generation.

Next, download all the files from Github, `https://github.com/sys-bio/aritificial_random_signaling_network`, into one folder. At the Julia console type:

`include("pathto\\Ground_truth_generation.jl")` to run the main script.

Note that `pathto` is the path you where you saved the network generation scripts to, and the double backslash is to avoid the Julia misinterpreting the backslash as a control character. A random signaling network will be generated in SBML [21] format called `sampleNetwork.xml`. There are some configuration settings in the Julia script file `Ground_truth_generation.jl`. These include:

1. The number of species, `nSpeces`, and maximum number of reactions `nRxns_limitation`.
2. Randomly assigned ranges for species concentrations and rate constants can be modified via `rnd_species` and `rnd_parameter`.
3. `concentration_perturb` can be used to set the factor that perturbs the concentration at the input species. Default is set to `2`.
4. The number of random networks to generate can be set by changing the variable `sampleSize`.

The file `roadrunner_c_api.dll` is the dynamic library for libRoadRunner that is used to provide SBML simulation support [17].

References

[1] Bezanson, Jeff & Edelman, Alan & Karpinski, Stefan & Shah, Viral. (2015). Julia: A Fresh Approach to Numerical Computing. 10.1137/141000671.

[2] Logue, Jeremy & Morrison, Deborah. (2012). Complexity in the signaling network: Insights from the use of targeted inhibitors in cancer therapy. Genes & development. 26. 641-50. 10.1101/gad.186965.112.

[3] Kuperstein, Inna & Bonnet, E & Nguyen, Hien-Anh & Cohen, David & Viara, Eric & Grieco, Luca & Fourquet, S & Calzone, Laurence & Russo, Christophe & Kondratova, Maria & Dutreix, Marie & Barillot, Emmanuel & Zinovyev, Andrei. (2015). Atlas of Cancer Signalling Network: A systems biology resource for integrative analysis of cancer data with Google Maps. Oncogenesis. 4. 10.1038/oncsis.2015.19.

[4] Jaqaman, Khuloud & Danuser, Gaudenz. (2006). Linking data to models: Data regression. Nature reviews. Molecular cell biology. 7. 813-9. 10.1038/nrm2030.

[5] Lillacci, Gabriele & Khammash, Mustafa. (2010). Parameter Estimation and Model Selection in Computational Biology. PLoS computational biology. 6. e1000696. 10.1371/journal.pcbi.1000696.

[6] Hengl, Stefan & Kreutz, Clemens & Timmer, J & Maiwald, Thomas. (2007). Data-Based Identifiability Analysis of Nonlinear Dynamical Models. Bioinformatics (Oxford, England). 23. 2612-8. 10.1093/bioinformatics/btm382.

[7] Oana, Chis & Banga, Julio & Balsa-Canto, Eva. (2011). Structural Identifiability of Systems Biology Models: A Critical Comparison of Methods. PloS one. 6. e27755. 10.1371/journal.pone.0027755.

[8] Bauer, Peter & Thorpe, Alan & Brunet, Gilbert. (2015). The quiet revolution of numerical weather prediction. Nature. 525. 47-55. 10.1038/nature14956.

[9] Pierce, David & Barnett, Tim & Santer, Benjamin & Gleckler, Peter. (2009). Selecting Global Climate Models for Regional Climate Change

208 Studies. Proceedings of the National Academy of Sciences of the United
209 States of America. 106. 8441-6. 10.1073/pnas.0900094106.

210 [10] Flato, Gregory & Marotzke, J. & Abiodun, Babatunde & Braconnot,
211 Pascale & Chou, Sin Chan & Collins, W. & Cox, Peter & Driouech,
212 Fatima & Emori, S. & Eyring, V. & Forest, Chris & Gleckler, P. &
213 Guilyardi, Eric & Jakob, C. & Kattsov, V. & Reason, Chris & Rum-
214 mukainen, M.. (2013). Evaluation of climate models.

215 [11] Seneta, Eugene. (2004). Fitting the variance-gamma model to
216 financial data. Journal of Applied Probability. 41. 177-187.
217 10.1017/S0021900200112288.

218 [12] Alexander, R. & Clarke, C. & Pringle, J.. (2006). Photoevapora-
219 tion of protoplanetary discs – I. Hydrodynamic models. MNRAS. 369.
220 10.1111/j.1365-2966.2006.10293.x.

221 [13] Basson, M.. (2012). Signaling in Cell Differentiation and Morphogen-
222 esis. Cold Spring Harbor perspectives in biology. 4. 10.1101/cshper-
223 spect.a008151.

224 [14] Klipp, Edda & Liebermeister, Wolfram. (2006). Klipp, E. & Lieber-
225 meister, W. Mathematical modeling of intracellular signaling pathways.
226 BMC Neurosci. 7, S10. BMC neuroscience. 7 Suppl 1. S10. 10.1186/1471-
227 2202-7-S1-S10.

228 [15] Somogyi, Endre & Bouteiller, Jean-Marie & Glazier, James & König,
229 Matthias & Medley, Kyle & Swat, Maciej & Sauro, Herbert. (2015).
230 libRoadRunner: A High Performance SBML Simulation and Analysis
231 Library. Bioinformatics. 31. 10.1093/bioinformatics/btv363.

232 [16] Sever, Richard & Brugge, Joan. (2015). Signal Transduction in Can-
233 cer. Cold Spring Harbor Perspectives in Medicine. 5. a006098-a006098.
234 10.1101/cshperspect.a006098.

235 [17] Choi, Kiri & Hellerstein, Joesph & Wiley, H Steven & Sauro Herbert M.
236 (2018). Inferring Reaction Networks using Perturbation Data. BioRxiv.
237 1-9. 10.1101/351767.

238 [18] Somogyi, Endre & Bouteiller, Jean-Marie & Glazier, James & König,
239 Matthias & Medley, Kyle & Swat, Maciej & Sauro, Herbert. (2015).

libRoadRunner: A High Performance SBML Simulation and Analysis Library. Bioinformatics. 31. 10.1093/bioinformatics/btv363.

[19] Markevich, Nikolai & Hoek, Jan & Kholodenko, Boris. (2004). Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. The Journal of cell biology. 164. 353-9. 10.1083/jcb.200308060.

[20] Klinke, David J. (2010). Signal Transduction Networks in Cancer: Quantitative Parameters Influence Network Topology. Cancer Research. 70. 1773-82. 10.1158/0008-5472.CAN-09-3234.

[21] Hucka, Michael & Finney, A. & Sauro, Herbert & Bolouri, H. & Doyle, John & Kitano, Hiroaki & Arkin, Adam & Bornstein, B.J. & Bray, D. & Cornish-Bowden, Athel & Cuellar, A.A. & Dronov, S. & Gilles, E.D. & Ginkel, M. & Gor, V. & Goryanin, Igor & Hedley, W.J. & Hodgman, Charlie & Hofmeyr, Jan-Hendrik & Wang, J.. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics. 19. 524-531. 10.1093/bioinformatics/btg015.