Highlights

Generating synthetic signaling networks for in silico modeling studies

Jin Xu,H Steven Wiley,Herbert M Sauro

- We provide a Julia script to generate synthetic random signaling networks.
- Our generated synthetic random signaling networks had high topological similarities with signaling networks from the BioModels Database.
- The behavioural dynamics of our synthetic signaling networks are similar to the dynamical behaviors of signaling networks from the BioModels Database.

Generating synthetic signaling networks for in silico modeling studies

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

[a]Department of Bioengineering, University of Washington, Seattle 98195, WA, USA
[b]Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland 99352, WA, USA

ARTICLE INFO

Keywords:
synthetic models
reaction-based models
biochemical networks
systems biology

ABSTRACT

Predictive models of signaling pathways have proven to be difficult to develop. The reasons for this include the uncertainty in the number of species, the complexity of 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. 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 the performance of any new algorithm. To meet this need, we developed a method for generating artificial signaling networks that are reasonably realistic and thus could be treated as ground truth models. These synthetic models could then be used to generate synthetic data for developing and testing algorithms designed to recover the underlying network topology and associated parameters. To determine whether our generated signaling networks display meaningful behavior, we compared them with signaling networks from the BioModels Database. This comparison indicated that our generated signaling networks had high topological similarities with BioModel signaling networks with respect to the reaction degree and distance distributions. In addition, our synthetic signaling networks had similar behavioral dynamics with respect to both steady states and oscillations, suggesting that our method generated synthetic signaling networks comparable with BioModel ones and thus could be useful for building network evaluation tools.

1. Introduction

In this article, we described a Julia [1] script that can be used to generate synthetic signaling networks. Such networks could 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 sixty 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 of 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 signaling networks that are reasonably realistic and 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 described 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 before 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

*Corresponding author
jin.xu.phys@gmail.com (J. Xu)
ORCID(s): 0000-0001-6738-9979 (J. Xu); 0000-0003-0232-6867 (H.S. Wiley); 0000-0002-3659-6817 (H.M. Sauro)

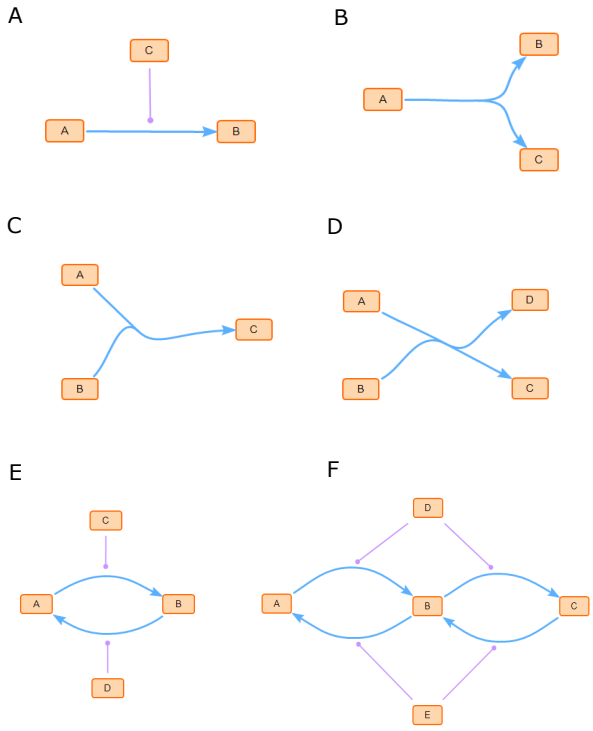

Figure 1: The synthetic random networks were composed of six types of reactions: (A) catalyzed uni-uni, (B, C) binding and unbinding reactions, (D) exchange reactions, (E) catalyzed phosphorylation-dephosphorylation, and (F) dual phosphorylation-dephosphorylation cycles. Figure was generated by SBcoyote [17].

networks contribute to the disease would 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. 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 could appropriately resemble natural ones.

2. Methods

To generate synthetic networks, we first created a list of unit processes that are to be found in most signaling networks. These were shown in Figure 1, including catalyzed transformations (A), three binding and unbinding reactions (B-D), and two phosphorylation/dephosphorylation units (E, F) which included single (E) and doubly-phosphorylated (F) motifs. The three binding and unbinding reactions (B-D) followed mass-action kinetic rate laws. The other three units (A, E, F) followed reversible Michaelis-Menten rate laws. Their corresponding rate equations are shown in Table 1. All rate constants and species concentrations were assigned randomly. The single-cycle case used two reversible Michaelis-Menten kinetics for both phosphorylation and dephosphorylation. For the double circle case, we made use of four reversible Michaelis-Menten kinetics, which was a special case of the single cycle.

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

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)	$v_{F1} = \frac{D(k_{f1} \cdot A/K_{A1} - k_{r1} \cdot B/K_{B1})}{1+A/K_{A1}+B/K_{B1}}$ $v_{F2} = \frac{D(k_{f2} \cdot B/K_{B2} - k_{r2} \cdot C/K_{C2})}{1+B/K_{B2}+C/K_{C2}}$ $v_{F3} = \frac{E(k_{f3} \cdot C/K_{C3} - k_{r3} \cdot B/K_{B3})}{1+C/K_{C3}+B/K_{B3}}$ $v_{F4} = \frac{E(k_{f4} \cdot B/K_{B4} - k_{r4} \cdot A/K_{A4})}{1+B/K_{B4}+A/K_{A4}}$

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

However, not all initially generated networks are viable or useful. Thus, the algorithm imposed structural constraints that must be followed. These included requiring that all species must be connected to a path in the network that connected the input to the output species, which ensured that no isolated species existed. In addition, this prevented reaction fragments from being isolated from the main body.

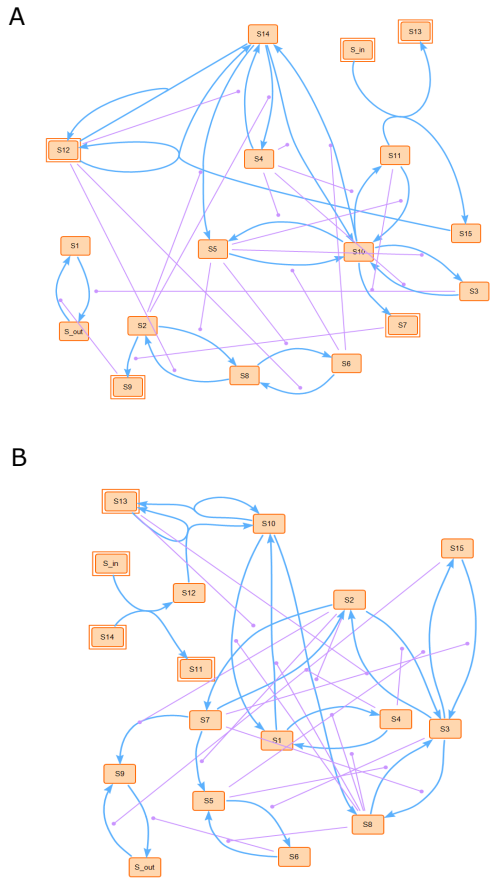

Figure 2: Random signaling networks with 15 species in addition to an input and output species (`S_in` and `S_out`) and 22 catalytic interactions are shown in blue lines with a circle at the end. Figure was generated by SBcoyote [17].

There were also dynamic constraints in the network generation system. Networks that could not reach a steady state were excluded. We also found that, although some networks appeared complex and connected, perturbations to the input failed to propagate to the output. Because such models could not process information, they were also excluded. We also excluded networks where the input species was directly connected to the output species. We required that there be at least three reactions between the input and output species.

The code to generate the synthetic networks was written using the Julia Language (`https://julialang.org/`) and the simulations were done using the libRoadRunner library [15, 18]; so that we could also export the result as a file with the Systems Biology Markup Language (SBML) format [19]. See Appendix for details.

3. Results

Figure 2 illustrates two examples of randomly generated signaling networks. Figure 2A included five types of chemical reaction processes in Table 1. Species were labelled `S1` to `S15`. These include three uni-uni processes: `S2` → `S9` catalyzed by `S7`, `S14` → `S5` catalyzed by `s2`, and `S10` → `S7` catalyzed by `S11`; one bi-uni reaction: `S14 + S12` → `S12`; two bi-bi reactions: `S11 + S_in` → `S15 + S13` and `S12 + S15` → `S12 + S14`; two single phosphorylation-dephosphorylation cycles: `S1` ⇋ `S_out` catalyzed by `S3` and `S9` and `s4` ⇋ `S14` catalyzed by `S2` and `S12`; and three dual phosphorylation-dephosphorylation cycles `S2` ⇋ `S8` ⇋ `s6` catalyzed by `S5` and `S12`, `S3` ⇋ `S10` ⇋ `S11` catalyzed by `S4` and `S5`, `S5` ⇋ `S10` ⇋ `s14` catalyzed by `S4` and `S6`. Figure 2B included the 6th type of chemical reaction process, namely one uni-bi reaction: `S10` → `S10 + S13`. As shown, all the random signaling networks had 15 species in addition to input and output species, with 22 reactions. The time taken to generate a single network that satisfied the constraints described in the Methods section could range from several minutes to several hours for large networks (Table 2).

Table 2
The time taken to generate and find a qualified random signaling network depended on the size of the network. The size of the random network was represented by the number of species involved (input and output species were not included). The errors represented the standard deviations from ten independent runs. All computations reported were done using an Intel i7 9700 processor running at 3.00 GHz with 32GB RAM using Windows 10.

Species #	Reactions #	Time (sec)
15	22	265±316
20	29	3447±2925
25	35	6917±7547

4. Discussion

To illustrate the topologies and dynamics of our generated synthetic random signaling networks, we compared them with signaling network models found in the BioModels Database (https://www.ebi.ac.uk/biomodels/). Using the BioModels Database we searched using the category of "Individual Models" and the keywords "signal(l)ing networks". There were 88 "Manually Curated" signaling models available to download. We downloaded the 88 models and have made them available on GitHub (`https://github.com/sys-bio/artificial_random_signaling_network`). We analyzed these models and found the average species number per model was 27±2 and the average number of reactions per model was 35±3. Therefore, we generated 88 random synthetic signaling networks with 27 species and 35 reactions to compare. The generated 88 synthetic networks are available on GitHub (`https://github.com/sys-bio/artificial_random_signaling_network`).

There were two primary properties of the models: topology and dynamics. From the perspective of topology, we first compared the distribution of reaction degree of species, defined as the number of reactions one species belonged to. We also observed the distribution of reaction distance between a random pair of species, defined as the minimum number of reactions between the two species. As an extreme case, if two

species belonged to one reaction, the reaction distance was zero. Figure 3, shows the two topological distributions of our generated synthetic random signaling networks compared with the signaling networks from the BioModels Database. To quantitatively compare the two topological distributions, we calculated the Bhattacharyya Distance [20] between the distributions of our generated signaling networks and signaling networks from BioModels. In general, our synthetic signaling networks were comparable with the BioModels signaling networks with Bhattacharyya Distances at approximately 0.06 and 0.02 for the reaction degree and reaction distance distributions respectively. From the perspective of dynamics, we looked at the steady state and oscillatory behavior, shown in Figure 4. Our artificial random signaling networks could generate comparable steady state and oscillatory behavior, which were also commonly observed in the signaling networks from the BioModels Database. Among the 88 signaling networks from the BioModels Database, 56 models showed steady states, and nine models that showed dynamics with oscillations. While, among our generated random signaling networks, there was only one case with oscillations.

There are two existing synthetic random reaction network generating tools, i.e., SMGen [21] and SBbadger [22]. They focus on different perspectives, however, none of the two tools considered the phosphorylation and dephosphorylation cycles which are the critical units in signaling networks. See the comparison in Table 3.

Table 3
Comparison with two other synthetic reaction network tools.

Major differences	Our method	SMGen	SBbadger
Programming language	Julia	Python	Python
Interface	script	terminal	GUI
(De)phosphorylation cycles	yes	no	no
Species counts	definable	definable	Max
reaction counts	definable	Min	Min

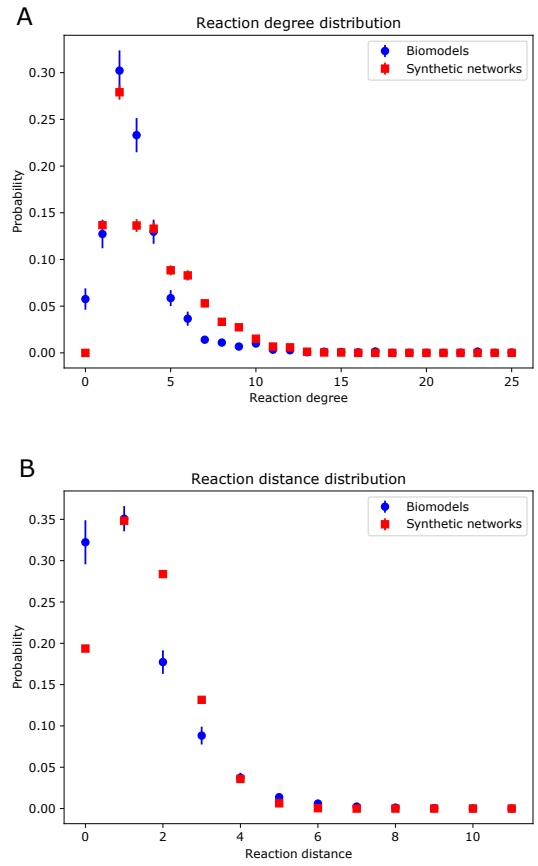

Figure 3: The reaction degree distributions (A) and reaction distance distributions (B) of two sets of signaling network models. The legend "Biomodels" represented the downloaded signaling networks from the BioModel Database. The legend "Synthetic networks" meant the models generated by our method.

5. Conclusions and limitations

The Julia programming language has gained traction with the systems biology community in recent years [23]. Therefore, we developed a Julia script that could allow users to computationally generate artificial random signaling networks to test novel algorithms for parameter fitting, topology mapping, and other analyses. To illustrate whether our generated signaling networks display meaningful behavior, we compared them with signaling networks from the BioModels Database. The two sets of networks had high similarity from the perspective of topology, namely their reaction degree distributions and reaction distance distributions were similar with low Bhattacharyya Distance. In addition, our artificial random signaling networks could generate comparable steady states and oscillations, which were also commonly observed in the signaling networks from the BioModels Database. As a comparison with the other two existing tools, SMGen [21] and SBbadger [22], our method considered the critical units of phosphorylation and dephosphorylation cycles for signaling networks. Additionally, users could define both species and reaction counts within our Julia script.

Limitations

The current approach we took has some limitations. The first is that the generated networks did not include any kind of sequestration in the phosphorylation and dephosphorylation cycles. Given the comparable concentrations of substrate and catalytic proteins found in signaling phosphatases, it is now recognized that sequestration [24, 25] can have a significant effect on signaling network behavior [26, 27]. Such effects can't be modeled with the current version of the software. This would be something to consider for future work.

Another aspect is that we did not examine whether our generated networks had similar graph metrics to networks found in nature. Currently, we only compared our networks to network models found in the BioModels Database. A

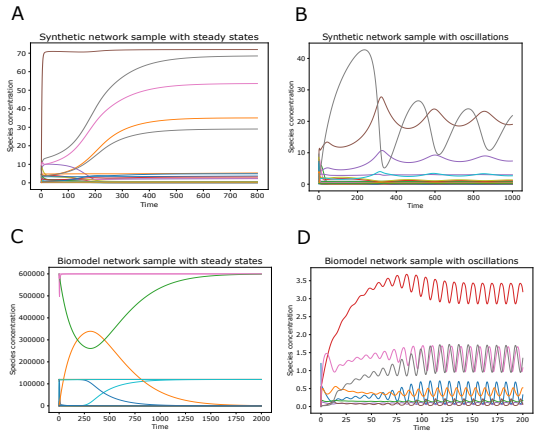

Figure 4: Time-course simulations for four models of signaling networks. The figures showed how the concentrations of species reached their steady state. A and B represented the steady states and oscillations from two sample models of our generated synthetic random signaling networks made up of 27 species (including the input and output species) and 35 reactions. C and D showed the steady states and oscillations from two sample models of signaling networks downloaded from the BioModels Database. C and D showed the dynamic behaviors of BIOMD0000000427 and BIOMD0000000720.

more thorough investigation would need to be carried out to investigate how the dynamic behavior compares to natural networks. We also did not impose biophysical constraints on the kinetic parameters. Because network behavior is a function of both topology and species parameters [28], such constraints could reduce the number of viable topological configurations. Finally, further analysis of which constraints must be either imposed or relaxed to generate realistic network behavior could reveal important principles that real biological networks follow during their evolution.

6. Acknowledgements

This work was financially supported by the National Institute of General Medical Sciences [grant number GM12303201] and the National Cancer Institute [grant number U01CA227544]. 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 the development of some of the early code.

7. Declaration of competing interest

None.

A. Code availability

The Julia script for generating the synthetic signaling networks is called Ground_truth_generation.jl and is available on GitHub at (`https://github.com/sys-bio/artificial_random_signaling_network`). This Julia script was implemented in Julia 1.6 on Windows 10. To use the Julia script, first install Julia by going to the Julia website (`https://julialang.org/`). Once you have the Julia console open, make sure you have `RoadRunner` installed by typing `import Pkg` followed by `Pkg.add("RoadRunner")`. RoadRunner.jl is a Julia package for the library of libRoadRunner that is used to provide SBML simulation support [18]. Follow the Julia warnings to install other necessary Julia packages if they have not been installed already.

Next, download the script from GitHub (`https://github.com/sys-bio/artificial_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` was the path where you saved the network generation script, and the double backslash was to avoid Julia misinterpreting the backslash as a control character. A random signaling network would be generated in SBML format called `sampleNetwork0.xml` before the steady state and `sampleNetwork.xml` after the steady state. There are some configuration settings in the Julia script file `Ground_truth_generation.jl`. These included:

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

CRediT authorship contribution statement

Jin Xu: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing-original draft. **H Steven Wiley:** Funding acquisition, Writing-review and editing. **Herbert M Sauro:** Conceptualization, Funding acquisition, Writing-review and editing.

References

[1] J. Bezanson, A. Edelman, S. Karpinski, V. B. Shah, Julia: A fresh approach to numerical computing, SIAM Review 59 (2017) 65–98.

[2] J. Logue, D. Morrison, Complexity in the signaling network: Insights from the use of targeted inhibitors in cancer therapy, Genes and development 26 (2012) 641–50.

[3] I. Kuperstein, E. Bonnet, H.-A. Nguyen, D. Cohen, E. Viara, L. Grieco, S. Fourquet, L. Calzone, C. Russo, M. Kondratova, M. Dutreix, E. Barillot, A. Zinovyev, Atlas of cancer signalling network: A systems biology resource for integrative analysis of cancer data with Google Maps, Oncogenesis 4 (2015).

[4] K. Jaqaman, G. Danuser, Linking data to models: Data regression, Nature reviews. Molecular cell biology 7 (2006) 813–9.

[5] G. Lillacci, M. Khammash, Parameter estimation and model selection in computational biology, PLoS computational biology 6 (2010) e1000696.

[6] S. Hengl, C. Kreutz, J. Timmer, T. Maiwald, Data-based identifiability analysis of nonlinear dynamical models, Bioinformatics (Oxford, England) 23 (2007) 2612–8.

[7] C. Oana, J. Banga, E. Balsa-Canto, Structural identifiability of systems biology models: A critical comparison of methods, PloS one 6 (2011) e27755.

[8] P. Bauer, A. Thorpe, G. Brunet, The quiet revolution of numerical weather prediction, Nature 525 (2015) 47–55.

[9] D. Pierce, T. Barnett, B. Santer, P. Gleckler, Selecting global climate models for regional climate change studies, Proceedings of the National Academy of Sciences of the United States of America 106 (2009) 8441–6.

[10] G. Flato, J. Marotzke, B. Abiodun, P. Braconnot, S. C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason, M. Rummukainen, Evaluation of climate models, Cambridge University Press, Cambridge, UK, 2013, pp. 741–882. doi:10.1017/CBO9781107415324.020.

[11] E. Seneta, Fitting the variance-gamma model to financial data, Journal of Applied Probability 41 (2004) 177–187.

[12] R. Alexander, C. Clarke, J. Pringle, Photoevaporation of protoplanetary discs – I. Hydrodynamic models, MNRAS 369 (2006).

[13] M. Basson, Signaling in cell differentiation and morphogenesis, Cold Spring Harbor perspectives in biology 4 (2012).

[14] E. Klipp, W. Liebermeister, Klipp, E. and Liebermeister, W. Mathematical modeling of intracellular signaling pathways. BMC Neurosci. 7, S10, BMC neuroscience 7 Suppl 1 (2006) S10.

[15] E. Somogyi, J.-M. Bouteiller, J. Glazier, M. König, K. Medley, M. Swat, H. Sauro, libRoadRunner: A high performance sbml simulation and analysis library, Bioinformatics 31 (2015).

[16] R. Sever, J. Brugge, Signal transduction in cancer, Cold Spring Harbor Perspectives in Medicine 5 (2015) a006098–a006098.

[17] J. Xu, G. Geng, N. D. Nguyen, C. Perena-Cortes, C. Samuels, H. M. Sauro, SBcoyote: An extensible python-based reaction editor and viewer, Biosystems 232 (2023) 105001.

[18] C. Welsh, J. Xu, L. Smith, M. König, K. Choi, H. Sauro, libRoadRunner 2.0: A high performance sbml simulation and analysis library, Bioinformatics 39 (2022).

[19] S. Keating, D. Waltemath, M. König, F. Zhang, A. Dräger, C. Chaouiya, F. Bergmann, A. Finney, C. Gillespie, T. Helikar, S. Hoops, R. Malik Sheriff, S. Moodie, I. Moraru, C. Myers, A. Naldi, B. Olivier, S. Sahle, J. Schaff, J. Zucker, SBML level 3: an extensible format for the exchange and reuse of biological models, Molecular Systems Biology 16 (2020).

[20] A. Bhattacharyya, On a measure of divergence between two statistical populations defined by their probability distributions., Bulletin of the Calcutta Mathematical Society 35 (1943) 99–109.

[21] S. G. Riva, P. Cazzaniga, M. S. Nobile, S. Spolaor, L. Rundo, D. Besozzi, A. Tangherloni, SMGen: A generator of synthetic models of biochemical reaction networks, Symmetry 14 (2022).

[22] M. A. Kochen, H. S. Wiley, S. Feng, H. M. Sauro, SBbadger: biochemical reaction networks with definable degree distributions, Bioinformatics 38 (2022) 5064–5072.

[23] E. Roesch, J. G. Greener, A. L. MacLean, H. Nassar, C. Rackauckas, T. E. Holy, M. P. Stumpf, Julia for biologists, Nature Methods 20 (2023) 655–664.

[24] N. Markevich, J. Hoek, B. Kholodenko, Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades, The Journal of cell biology 164 (2004) 353–9.

[25] A. Ciliberto, F. Capuani, J. J. Tyson, Modeling networks of coupled enzymatic reactions using the total quasi-steady state approximation, PLoS computational biology 3 (2007) e45.

[26] N. Blüthgen, Sequestration shapes the response of signal transduction cascades, IUBMB life 58 (2006) 659–663.

[27] C. Gomez-Uribe, G. C. Verghese, L. A. Mirny, Operating regimes of signaling cycles: statics, dynamics, and noise filtering, PLoS computational biology 3 (2007) e246.

[28] D. Klinke, Signal transduction networks in cancer: Quantitative parameters influence network topology, Cancer research 70 (2010) 1773–82.