In Silico Experimentation on the Use of Signal Flow Analysis

Madeleine S Gastonguay[a], *Lauren Marazzi*[a], *Paola Vera-Licona*[a]

[a] *Center for Quantitative Medicine, UConn Health, Farmington, CT 06030*

INTRODUCTION

Intracellular signaling networks can represent large-scale signaling processes that occur within cells. Represented as a graph, the nodes of an intracellular signaling network represent molecular components of a cell and edges are the regulatory interactions in between them. Unlike mathematical models, which require knowledge of system dynamics such as kinetic parameters or logic formalisms, signaling networks can be constructed when this information is unknown or difficult to determine. Though signaling networks do not describe the temporal behavior of a system, dynamical systems based approaches for network analysis have been developed. One such approach is the estimation of network dynamics based on network structure. Lee and Cho developed a method to estimate signal flow based on the network topology and predict network steady states from user provided initial conditions. In this analysis, we determine the appropriate uses and limitations of this signal flow analysis (SFA) algorithm [1].

We took two approaches to answering this question:

1. Determine the ability of SFA to estimate attractors of a boolean model. The results of this study will allow us to ascertain if we can estimate the attractor landscape using SFA.
2. Identify appropriate clustering methods to approximate attractor landscape and classify attractors of both the Boolean model and the signaling network.

Analyses were done on a signaling network for T-LGL Leukemia constructed by Zhang et al. [2]. This network was also translated into a Boolean model (**Figure 1**) [2]. The network consists of 60 nodes - including "Cytoskeleton Signaling", "Proliferation", and "Apoptosis" as markers of cell fate - and 142 edges. Network reduction techniques [3] and control target identification methods [4] have previously been applied to this network.

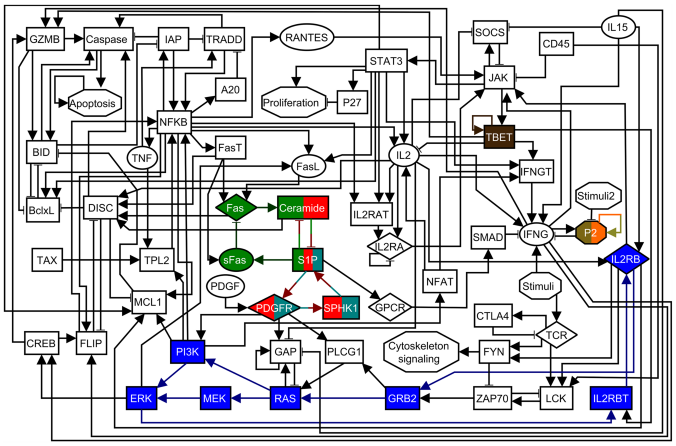


Figure 1. Boolean Network for T-LGL Leukemia. Colored nodes were identified as nodes that can drive the system to desired cell fates.

SOFTWARE AND ALGORITHMS

We are using the Signal Propagation (SP) algorithm developed by Cho et al. to estimate the attractors of the signaling network based solely on the network topology via signal flow between nodes [1]. This algorithm solves for the state of node x_i at time $t+1$ by

$$x_i(t+1) = \alpha \Sigma_j W_{ij} x_j(t) + (1-\alpha) b_i \quad \textbf{Equation 1.}$$

where α is a hyperparameter, b is the basal (initial) activity of node x_i, and x_j is the state of an incoming node to node x_i. The link weight matrix W contains normalized values for each edge W_{ij} using

$$W_{ij} = \frac{sign(ij)}{\sqrt{(D_{out})_j (D_{in})_i}} \quad \textbf{Equation 2.}$$

where $sign(ij)$ is the sign of the edge (1 for activating edges, -1 for inactivating edges), D_{out} is the out-degree of source node j, and D_{in} is the in-degree of target node i. This equation is solved until the difference of between $x(t+1)$ and $x(t)$ is less than a tolerance threshold (10^{-5}).

The output of SFA is the log steady state activity of each node. The magnitude of this output cannot be interpreted on its own. Instead, the sign of the direction of activity change (DAC) between two attractors must be considered [1]. Since the output are log values, this is effectively the log-fold change between two conditions.

To characterize the accuracy of SFA to estimate attractors of a signaling network, we identify the attractors of the corresponding Boolean model for comparison. To do so, we are using BoolNet, an open source R package that can identify attractors of the model and their basins [5]. Unlike the signaling network, we can identify the true attractors of the model without the need for estimation because we have the system dynamics as logic formalisms.

1 DETERMINE THE ABILITY OF SFA TO REPLICATE THE RESULTS OF A BOOLEAN MODEL

Replicate Zañudo and Albert's Analysis.

Zañudo and Albert used the T-LGL Leukemia Boolean model to validate their stable motif algorithm and identify interventions that drive the system to Leukemia or Apoptosis [4]. Previous work by Zhang et al. and Saadatpour et al. show that in the presence of IL15 and Stimuli signal, the network has three attractors: one associated to the apoptosis phenotype, and two associated to the Leukemia phenotype [2, 3]. These attractors were associated to the Apoptosis phenotype if the APOPTOSIS node was ON and the Leukemia phenotype if the APOPTOSIS node was OFF. Zañudo and Albert simulated the attractors of the network in the presence of IL15 and Stimuli and used their stable motif algorithm to identify stable motifs that can steer the model to either the Apoptosis associated attractor or the Leukemia associated attractors. To validate the intervention targets, they simulated the network with 100,000 random initializations and compared the probability that an arbitrary initial state leads to an apoptosis associated or leukemia associated phenotype with and without intervention of the stable motif nodes. In this step, IL15 and Stimuli were not initialized to ON, so the 100,000 initializations resulted in more than three attractors. They were associated to the Apoptosis or Leukemia phenotype based on the activity of the APOPTOSIS node. Zañudo and Albert found that the multi-node interventions resulted in the desired attractor being reached for 100% of the random initializations. Two of these interventions steered the system towards Leukemia associated attractors and six of the interventions steered the system towards the Apoptosis associated attractor (**Table 1**). While they only used the activity of the APOPTOSIS node to associate these randomly simulated attractors to phenotypes, they found common trends among several nodes of the Apoptosis and Leukemia associated attractors (**Table 2**). These 19 nodes can be used as marker nodes to associate randomly simulated attractors to ether the Leukemia or Apoptosis phenotype.

We were able to replicate these results in the Boolean model by fixing the intervention nodes ON (1) or OFF (0), and simulating with IL15 and Stimuli initialized as ON. All other network nodes were initialized as OFF. Oscillating at-

Table 1. Stable motif interventions identified by Zañudo and Albert that drive the Boolean model to Apoptosis or Leukemia when fixed in their designated orientation. The "Basin" size indicates how many states converge to the attractor resulting from fixing the nodes of the corresponding stable motif.

Cell Fate	Motif Label	Stable Motif Interventions	"Basin" Size
Leukemia	L1	Ceramide = OFF, SPHK1 = ON	19
	L2	Ceramide = OFF, PDGFR = ON	17
Apoptosis	A1	TBET = ON, Ceramide = ON, RAS = ON	10
	A2	TBET = ON, Ceramide = ON, GRB2 = ON	11
	A3	TBET = ON, Ceramide = ON, IL2RB = ON	13
	A4	TBET = ON, Ceramide = ON, IL2RBT = ON	14
	A5	TBET = ON, Ceramide = ON, ERK = ON	15
	A6	TBET = ON, Ceramide = ON, MEK = ON, PI3K = ON	10

Table 2. Nodes identified by Zañudo and Albert that exhibit consistent expression in Leukemia and Apoptosis associated attractors. Columns represent the two phenotypes and rows represent the activity of the readout nodes associated to the phenotype. Nodes in red indicate those that are ON in Leukemia associated attractors but OFF in Apoptosis associated attractors.

Activity	Leukemia Readout Nodes	Apoptosis Readout Nodes
ON	FasL, FasT, NFKB, TPL2, IFNGT, PDGFR, GPCR, S1P, SPHK1, PI3K	Apoptosis, Caspase, Ceramide, DISC, Fas, GZMB, BID
OFF	Apoptosis, Caspase, Ceramide, DISC, Fas, TRADD	IAP

tractors were condensed to quasi-attractors where the activity of oscillating nodes is denoted by an X as described by Zañudo and Albert [4]. The two Leukemia interventions (L1 and L2) resulted in two Leukemia associated attractors differing in the expression of P2, which is in a self loop. The six Apoptosis interventions (A1-A6) converged to the same attractor. These results agree with the findings of Zhang and Saadatpour [2, 3]. Although the six Apoptosis interventions resulted in the same attractor, each one has a different "basin" of attraction because the attractor was reached by fixing different nodes. **Table 1** includes the number of states that lead to the Apoptosis and Leukemia associated attractors when the corresponding stable motif nodes are fixed.

To replicate this analysis in the signaling network, we estimated the attractors of the same eight initial conditions and stable motif interventions on the signaling network using SFA. To fix the state of the intervention targets, we ran SFA iteratively and manually overrode their states after each time step until the algorithm converged to a steady state activity level. When we simulated the stable motif interventions in **Table 1** with SFA, eight separate attractors (two Leukemia associated and six Apoptosis associated) were reached (**Figure 2b**). It is difficult to compare the SFA output to the attractors of the Boolean network because the former is on a continuous scale while the latter is discrete. However, hierarchical clustering shows that the Apoptosis associated attractors cluster separately from the Leukemia associated attractors from both the signaling network and Boolean model (**Figure 2**).

Conclusions: These results show that SFA is producing more attractors than calculated from the Boolean model for the same initializations and fixed nodes. This may be because the SFA output is on a continuous scale while the Boolean attractors are discrete.

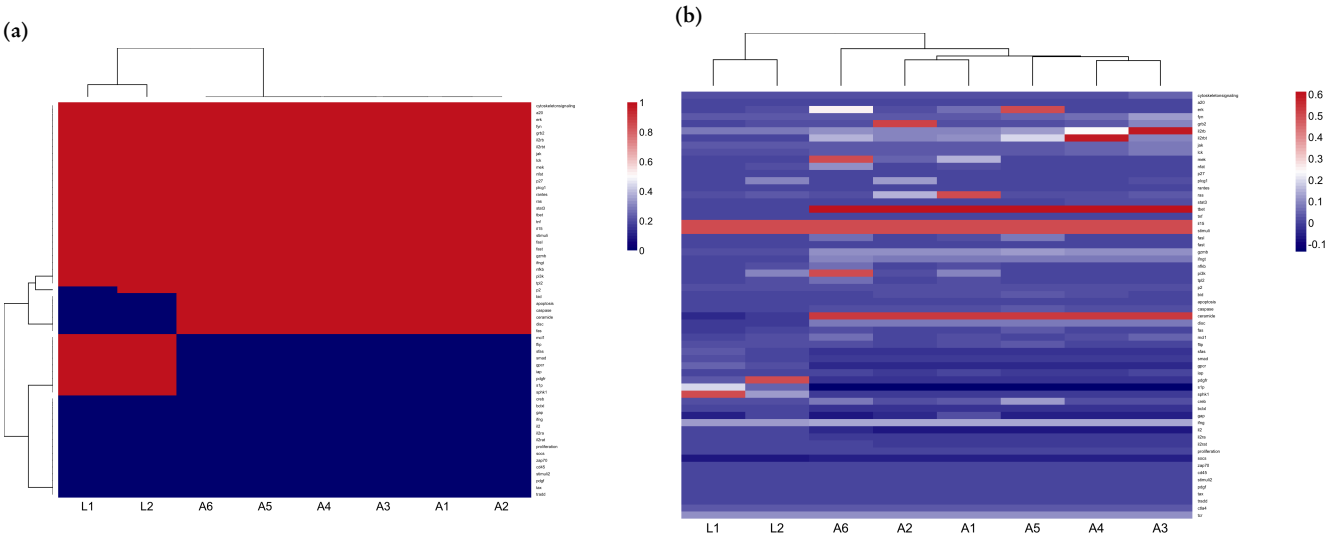


Figure 2. (a) Heatmaps with hierarchical clustering of attractors of the Boolean model simulated with fixed Stable Motif interventions leading to Apoptosis and Leukemia associated attractors. The two oscillating nodes, CTLA4 and TCR, were excluded from this heat map as they flip between ON and OFF in both leukemia and apoptosis associated attractors. (b) Heatmaps with hierarchical clustering of log steady state attractors of the signaling network estimated by SFA with fixed stable motif nodes. The color gradient indicates the z-score resulting from centering and standardizing the expression profile of each sample.

Optimal discretization of SFA output using DiscreeTest.

In order to compare the SFA output to the Boolean attractors, we applied multiple discretization methods to the SFA trajectory for the 8 initial states using GED protools [6]. The simulated SFA trajectory was treated as time series data, with rows representing individual genes and columns representing the log activity of each gene at each time step. All possible discretization methods that binarized the SFA output were applied to the rows. This included bidirectional kmeans discretization with 2 levels, equal width discretization with 2 levels, kmeans discretization with 2 levels, Max - Y% Max (for $Y = 25, 50, 75$), discretization through comparing to mean value, discretization through comparing to median value, equal frequency discretization with 2 levels (q2), target discretization threshold, and Top Y% discretizations (for $Y = 25, 50, 75$). Then, DiscreeTest was used to determine which discretization method worked the best [7]. First, DiscreeTest performs a sign test to compare the distribution of the discretized data to the original data. Next, DiscreeTest calculates the mean area between the curves of the original data and the discretized data. The method that minimizes this area is determined to be the optimal discretization method. This optimal method determined via DiscreeTest was not consistent for the SFA trajectories from the eight initial states (**Table 3**). In fact, three of the trajectories did not have any discretization method pass the sign test.

Table 3. Optimal Discretization Method for the SFA trajectory simulated with IL15 and Stimuli ON and Stable Motif nodes fixed as determined by DicsreeTest. Initializations with an Optimal Discretization Method "NONE" had no discretization methods pass the sign test.

Stable Motif Intervention	Optimal Discretization Method
L1	q2
L2	NONE
A1	NONE
A2	q2
A3	NONE
A4	q2
A5	top75
A6	top75

Conclusion: There is not an optimal discretization method that can be applied to all of the SFA trajectories.

Simulating Basins of Attraction.

Next, we wanted to see if we could replicate the basins of attraction of the Boolean model with SFA. **Table 1** shows the number of states that lead to each attractor when simulated with the specified intervention fixed in the Boolean model- the size of the "basin" of that attractor. We applied applied SFA to the network initialized with each of those states and fixed the corresponding intervention targets. While states in the same "basin" of an attractor of the Boolean model collapse to the same attractor, they lead to separate attractors when simulated from the network with SFA. This analysis produced 36 SFA attractors from initial states in the "basin" of Leukemia associated attractors of the Boolean model and 73 SFA attractors from initial states in the "basin" of the Apoptosis associated attractors of the Boolean model. We found that the SFA trajectory from any initial state typically reached a steady state in less time steps than the trajectory of the same state when simulated with the Boolean model. Although neither the trajectory nor the attractors produced by SFA appear to replicate that of the Boolean model, we wanted to see if the resultant attractors could be associated to the correct phenotype (ie. those initial states that lead to Leukemia (Apoptosis) associated attractors in the Boolean model lead to Leukemia (Apoptosis) associated attractors when simulated with SFA).

To associate the SFA attractors to biological phenotypes, we used the previously mentioned 19 nodes that Zañudo and Albert identified as markers of Leukemia and/or Apoptosis attractors from the Boolean model as readout nodes (RONs) (**Table 2**). It is important to note that a node need not be ON in one phenotype and OFF in the other. For example, IAP is OFF in Apoptosis associated attractors, but it can be ON or OFF in Leukemia associated attractors.

While the magnitude of the SFA output cannot be interpreted, we can use the log steady state value to calculate the direction of activity change (DAC) between two simulated attractors. We would expect the DAC between the RONs of an Apoptosis associated SFA attractor and a Leukemia associated SFA attractor to match the DAC of the RONs between the Apoptosis and Leukemia associated attractors of the Boolean model (**Table 4**). Thus, we computed the DAC between all possible pairs of SFA attractors resulting from states in the "basin" of the Apoptosis associated attractors and the states in the "basin" of the Leukemia associated attractors of the Boolean model. (DAC = Apoptosis associated attractor - Leukemia associated attractor). This left us with a total of $36 \cdot 73 = 2628$ comparisons. To reduce the emphasis on the magnitude of the DAC, the resultant DACs were grouped into 3 values: 1 for those greater than zero, 0 for those equal to zero, and -1 for those less than zero.

Apoptosis	BID	Caspase	Ceramide	DISC	Fas	FasL	FasT	GPCR	GZMB
IAP	IFNGT	NFKB	PDGFR	PI3K	S1P	SPHK1	TPL2	TRADD	

Table 4. DAC between RONs of Apoptosis and Leukemia attractors of the Boolean model (Apoptosis - Leukemia). Red indicates a positive DAC, blue indicates a negative DAC, and white indicates a DAC of 0.

The DAC of the RONs resulting from each of these Apoptosis-Leukemia pairs was compared to the DAC between the RONs of the Apoptosis and Leukemia attractors from the Boolean model by calculating the Hamming distance between the two vectors. We found that 4% of the SFA comparisons matched the DAC between the Boolean attractors in 10 out of the 19 RONs (52%), and 96% of the SFA comparisons matched the DAC between the Boolean attractors in 11 of the 19 RONs (58%). Upon further examination, we realized that the nodes that SFA predicted the incorrect DAC for were consistently RONs that had a DAC of zero between the Apoptosis and Leukemia Boolean attractors. When considering only those RONs that had a non-zero DAC, we found that SFA predictions of DAC greatly improved. 4% of the SFA comparisons matched the DAC between the Boolean attractors in 10 of the 11 (91%) of the readout nodes with a non-zero DAC, and the other 96% matched the DAC of 100% of these RONs. After a further exploration of the DAC between SFA predicted values of the RONs with a zero DAC in the Boolean model, there does not appear to be a pattern in the incorrectly predicted DAC.

Conclusions: We cannot replicate a DAC of zero between SFA attractors. This is potentially due to the number of sig-

nificant digits of the output since the SFA attractors are continuous while the Boolean attractors are discrete.

Tolerance Exploration.

We wanted to see if we could recover some of the zero DACs by altering the tolerance level for convergence of the SFA algorithm (default = 10^{-5}). **Table 5** shows that as we increased the tolerance level we are only able to capture at most 54% of the zeroes we should see for each RON with an approximately 12% decrease in the accuracy of predicting the DAC of the RONs with a non-zero DACs.

Table 5. Percent of comparisons with a DAC of zero for each of the 8 RONs that have a DAC of zero between the Apoptosis and Leukemia attractors from the Boolean Model (FasL, FasT, GZMB, NFKB, PI3K, TPL2, and TRADD) as the tolerance level for convergence of the SFA algorithm increases. Accuracy in predicting the DAC of other RONs is denoted by the percent of comparisons that predict a DAC of zero for RONs that should have a non-zero DAC. (ϵ = *machine epsilon*)

Tolerance Level	FasL	FasT	GZMB	IFNGT	NFKB	PI3K	TPL2	TRADD	Accuracy in other RONs
ϵ	0	0	0	0	0	0	0	0	100
10^{-12}	0	0	0	0	0	0	0	0	100
10^{-10}	0	0	0	0	0	0	0	0	100
10^{-8}	0	0	0	0	0	0	0	0	100
10^{-6}	0	0	0	0	0	0	0	0	100
10^{-5}	0	0.07	0	0	0	0	0	0	100
10^{-4}	0	0.29	0	0	0.43	0	0.07	0.22	99.93
10^{-3}	1.37	1.8	0	0	1.44	0.14	2.38	1.15	99.21
10^{-2}	21.34	53.57	0	4.04	10.53	15.36	26.6	31.22	88.46

Conclusions: Increasing the tolerance, thereby decreasing the number of significant figures, is not an effective way to recover DACs that should be zero. Thus, we are not able to use SFA to accurately reproduce DACs that are zero between the Boolean attractors and we decided to remove RONs with a DAC of zero.

Redefining Readout Nodes.

We are using readout nodes (RONs) as an internal control to classify the attractors produced by SFA. We started with the 19 nodes Zañudo and Albert observed to be in the same orientation for apoptosis and leukemia attractors of the Boolean model (**Table 2**). After our tolerance exploration, we removed eight nodes with a DAC of zero, leaving us with 11 RONs. After further literature search, we selected only those RONs that fit the following criteria:

1. Have a non-zero DAC between the Apoptosis and Leukemia attractors of the Boolean model
2. Are experimentally validated markers of Leukemia and/or Apoptosis
3. Are identified as markers of Apoptosis and Leukemia in Zañudo and Albert's work [4]
4. Are in the reduced network from Saadatpour et al. [3] (**Figure 3**)

This left us with seven RONs (**Table 6**). The only node that does not fit all 4 criteria is Inhibitors of Apoptosis (IAP), which is not in the reduced Saadatpour network. IAP is an experimentally validated as a regulator of Apoptosis and identified as a RON by Zañudo, so we thought is was a necessary to include [8].

Apoptosis	BID	Ceramide	DISC	Fas	IAP	S1P

Table 6. DAC between reduced RONs of Boolean Apoptosis and Leukemia attractors (Apoptosis - Leukemia). Red indicates a positive DAC and blue indicates a negative DAC.

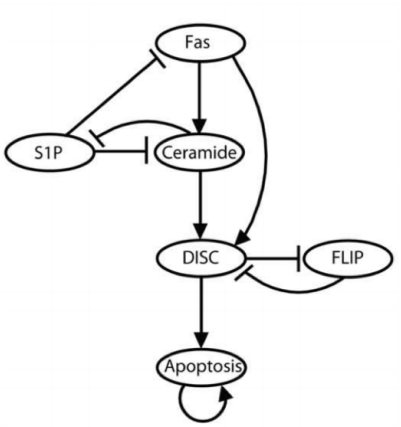


Figure 3. Resultant network from Saadatpour et al.'s network reduction method.

With the new readout nodes, 96% of the comparisons between SFA attractors from states in the "basin" of the Leukemia associated attractors and the Apoptosis associated attractors match the DAC between the Boolean attractors in all 7 RONs and 4% of the comparisons match 6 of the 7 RONs. Therefore, although the SFA attractors resulting from initializations in the "basin" of the Leukemia and Apoptosis attractors of the Boolean model do not converge to the same attractors as the Boolean model, they display the same trends in the DAC of the readout nodes as the Boolean Apoptosis and Leukemia associated attractors. Therefore, we can use associate the SFA attractors to a phenotype by calculating the DAC of the RONs.

Examining the log steady state activity predicted by SFA for the 36 attractors from states in the "basin" of the Leukemia associated attractors and the 84 attractors from states in the "basin" of the Apoptosis associated attractors provides an illustration of how these attractors differ. **Figure 4** compares the distribution of the raw log steady state SFA output for each of the readout nodes in apoptosis associated attractors and leukemia associated attractors. It is clear to see that the distribution of these values is different in Apoptosis associated and Leukemia associated attractors, making it possible to distinguish between the two phenotypes via the RONs of the attractors. Furthermore, the direction of activity change between the medians of these distributions for each of the RONs matches the expected DAC between the attractors of the Boolean model.

Conclusions: The results of the hamming distances between SFA and Booelan DACs indicate that although SFA cannot produce the same attractor that a Boolean model does, the attractor estimated from an initial condition by SFA can be associated to the same phenotype that the initial condition leads to in the Boolean model. The visualizations of the distribution of log steady state outputs for Leukemia and Apoptosis associated attractors estimated via SFA demonstrate that when stable motif nodes are fixed, different initial conditions leading to the Apoptosis associated attractor in the Boolean model lead to attractors with similar expression patterns of RONs when estimated with SFA, and these expression patterns are up-regulated or down-regulated compared to the RONs of attractors resulting from initial states in the basin of Leukemia associated attractors of the Boolean model. The DAC between the median expression values of the apoptosis and leukemia associated SFA attractors matches the DAC of the Boolean model. Hence, these attractors can be can be associated to phenotypes based on the DAC of their RONs.

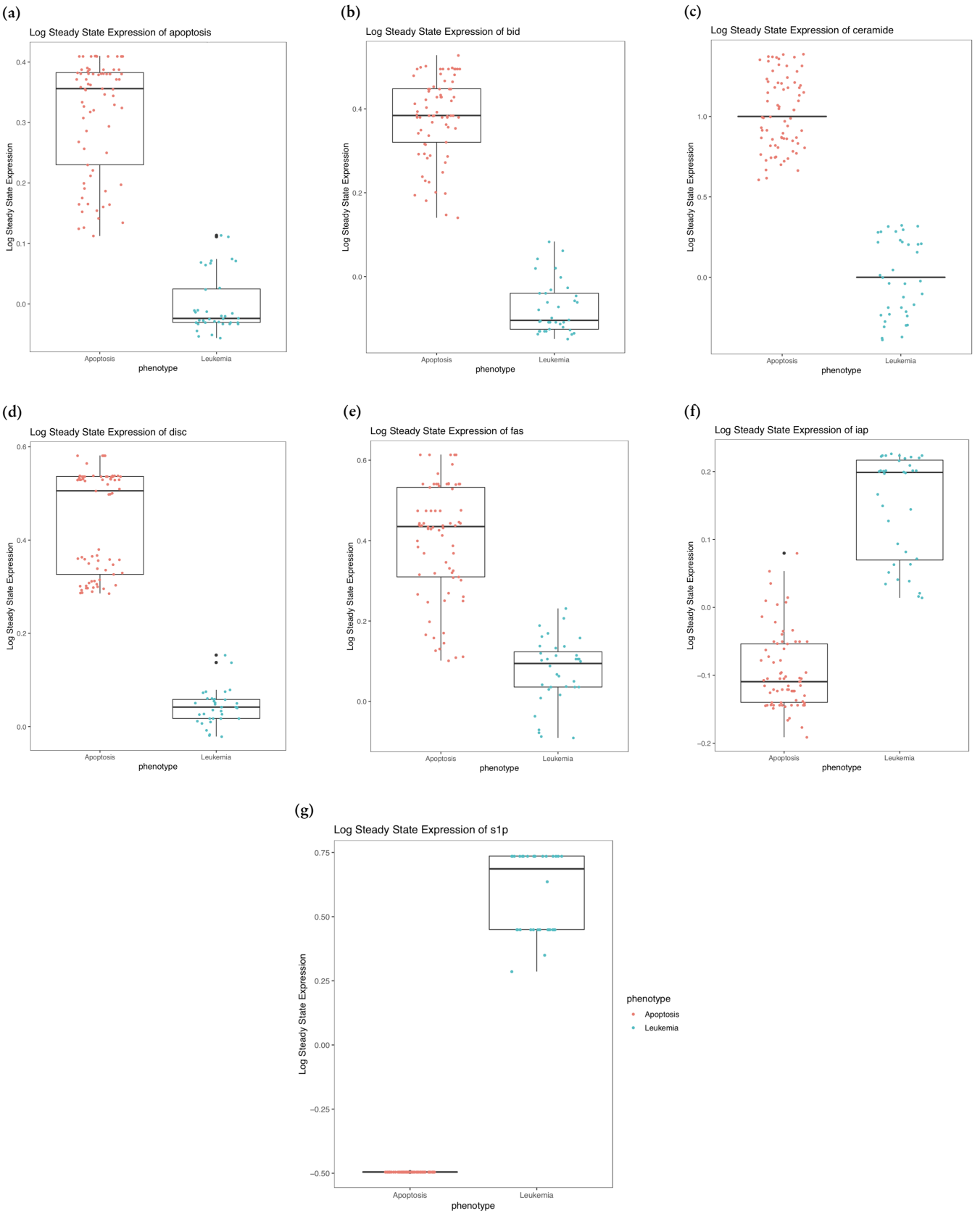


Figure 4. Boxplots of the log steady state value of RONs produced by SFA applied to initial conditions in the "basin" of leukemia and apoptosis associated attractors of the Boolean model. There is clearly a difference in the distribution of SFA values for Leukemia and Apoptosis associated attractors. Apoptosis, BID, Ceramide, and DISC, FAS are up-regulated in most Apoptosis associated attractors compared to Leukemia associated attractors. IAP and S1P are down-regulated in most apoptosis associated attractors compared to Leukemia associated attractors. These DACs match that of the Apoptosis and Leukemia associated attractors of the Boolean model, indicating that we can use the DAC of the RONs to classify them.

2 ANALYSIS OF CLUSTERING RANDOMLY SIMULATED SFA AND BOOLEAN ATTRACTORS.

Estimation of Attractor Landscape.

The current workflow we are using to estimate the Attractor Landscape of the signaling network is to 1) generate 100,000 random unique initial conditions, 2) simulate the resulting attractors with SFA, 3) cluster the SFA attractors using unsupervised k-means clustering, and 4) associate clusters to phenotypes based on where the reference attractors cluster. The reference attractors are SFA attractors resulting from initial states of experimental conditions corresponding to each phenotype of interest.

We wanted to test the ability of this workflow to associate attractors to phenotypes by applying it to the T-LGL signaling network. We generated 100,000 unique random initial conditions and simulated the resulting attractors with both the Boolean model and the static network. When simulating with the Boolean model, the initial states collapsed to 877 attractors. 31 of which were associated with Leukemia and 133 were associated with Apoptosis based on the activity level of the RONs (**Table 6**). The Apoptosis attractor and the two Leukemia attractors reached from simulating the Boolean model with Stimuli and IL15 ON and fixed stable motif targets identified by Zañudo and Albert (**Table 1**) were also reached from the 100,000 initial conditions with the unperturbed system (**Table 7**).

Table 7. Zañudo and Albert's stable motif interventions lead to one Apoptosis associated attractor and two Leukemia associated attractors which were also achieved by simulating the Boolean network without fixing stable motif nodes. The number of states in the basin of attraction of these attractors is shown in the above table, along with the number of states in those basins that are included in the 100,000 randomly simulated initial conditions.

Boolean Attractor	Basin Size From Unperturbed System	Number of states in the basin that are from our random 100,000 initial conditions
Apoptosis	9901	1396
Leukemia1	107	11
Leukemia2	615	82

According to the results from part 1, we would expect the initial conditions in the basins of these 3 attractors to lead to Leukemia or Apoptosis associated attractors when simulated with SFA. Just as we did in **Part 1**, for any of the 100,000 initial conditions that are in the basin of these three attractors, we simulated the resulting attractor with SFA on the unperturbed signaling network. Next, we compared the DAC between each possible comparison of attractors from states in the basin of Apoptosis and Leukemia associated attractors ($1396 \cdot (11 + 82) = 129828$ comparisons). We then computed the hamming distance between the DAC of the RONs of SFA attractors and the DAC of the RONs of the Boolean attractors, just as we did before (**Table 6**). The SFA attractors estimated from our random initial conditions exhibit a wide spread of accuracy in the predicted DAC of the seven RONs with a median value around 70% (**Figure 5**). This is contrary to what we observed in part 1 when we applied SFA to the static network with fixed intervention targets and initializations in the basin of the Leukemia and Apoptosis associated attractors.

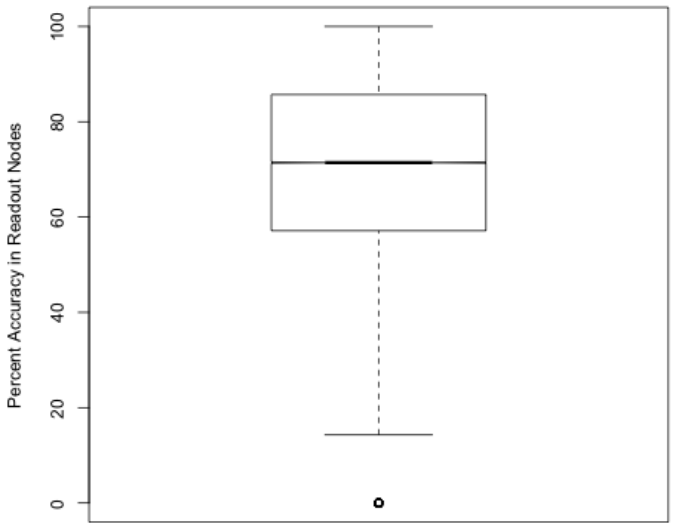


Figure 5. Distribution of hamming distances for the 129,828 Apoptosis - Leukemia comparisons. Percent accuracy denotes the percent of the seven RONs whose DAC between SFA attractors matches the DAC between the Boolean Apoptosis and Leukemia associated attractors.

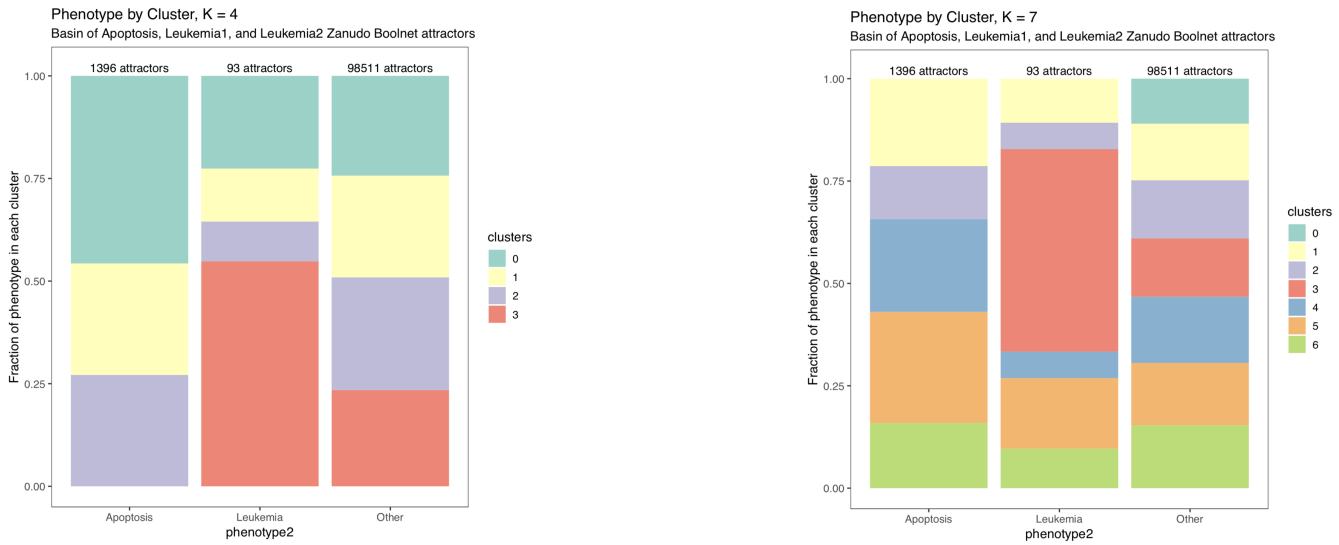
Analyzing Clustering Methods.

In order to identify which of the 100,000 random SFA attractors have similar characteristics, we use unsupervised k-means clustering to identify the "phenotype landscape" of the static network. This process was done on multiple manipulations of the SFA output. All three datasets were "discretized" as follows: positive values are represented as 1, negative values are -1, and zeros remain zero. The datasets are as follows (n is the number of network nodes: 60):

1. n-tuples of "discretized" log steady-state output from SFA
2. n-tuples of "discretized" DAC of each attractor compared to the SFA attractor estimated from the L1 stable motif
3. 2n-tuples of the concatenation of the above datasets

We ran k-means with k ranging from 2 to 15 in an attempt to identify the optimal k. An optimal k would have clustered the 1396 Apoptosis associated SFA attractors apart from the 93 Leukemia associated attractors (**Table 7**). However, we found that regardless of the number of centroids, the majority of Apoptosis associated attractors did not cluster separately from the majority of the Leukemia associated attractors. Thus, we could not distinctly associate each cluster to a phenotype.

In the previous clustering methods, we computed the DAC against an arbitrary Leukemia associated attractor. To limit the bias this introduced, we devised a new plan for k-means clustering. Instead of clustering the 2n-tuples of "discretized" log steady state and DAC against one leukemia associated attractor, we clustered with 94n-tuples: n observations of the "discretized" log steady state value for each network node, followed by the "discretized" DAC of that attractor against each of the 93 Leukemia associated attractors. When this dataset was clustered with four centroids, the majority of Leukemia associated attractors clustered together in cluster 3, which does not contain any Apoptosis associated attractors (**Figure 6a**). While the there was not one cluster in which the majority of Apoptosis associated attractors clustered, just less than 50% of them clustered in cluster 0 (**Figure 6a**). As the number of centroids increased, we consistently observed approximately 40% of the Leukemia associated attractors clustering together, but there was not one distinct Apoptosis associated cluster in which the majority of Apoptosis associated attractors cluster. See **Figure 6b** for an example with 7 centroids.



(a) K-means with 4 centroids. **(b) K-means with 7 centroids.**

Figure 6. Barplots of the proportion of Apoptosis associated attractors, Leukemia associated attractors, and attractors associated to other phenotypes that belong to each cluster with a) 4 centriods and b) 7 centroids. When k = 4, the majority of Leukemia attractors are in cluster 3, which has no Apoptosis clusters. Similarly, the majority of Apoptosis attractors are in cluster 0, which has fewer Leukemia attractors. When k = 7, the majority of Leukemia attractors are in cluster 3, but there is no clear Apoptosis cluster.

Seeing as approximately 40% of the Leukemia associated attractors cluster together for k-means with k of 4 through 12, the naturally question is whether or not this 40% consisted of the same attractors at each k value. We found that 28 of the 93 Leukemia associated attractors (30%) always clustered together, regardless of the k. Since the initial states these attractors were simulated from lead to Leukemia associated attractors in the Boolean model, and they consistently cluster together, we expect the attractors to exhibit similar DACs in their RONs. We computed the DAC for each of these 28 Leukemia associated attractors against the median log steady state value for all 93 Leukemia associated attractors. **Figure 7** shows that these 28 attractors do not have the same DAC for their RONs. Of note, the node APOPTOSIS has a positive DAC in 54% of the attractors, a negative DAC in 43% of the attractors, and a DAC of zero in 3% of the attractors. Since these are Leukemia associated attractors, APOPTOSIS should be downregulated in all of them. We see similar results for BID, DISC, and IAP, where about half of the attractors have a positive DAC and half have a negative DAC. Almost all of the 28 attractors exhibit the same DAC for Ceramide, FAS, and S1P.

One potential reason for the discrepancy in the DACs of these attractors is that the SFA output of these 28 attractors is being compared to the median of the 93 Leukemia associated attractors, so it makes sense that some attractors have values higher than the median (DAC of 1) and some to have values lower than the median (DAC of -1). Therefore, we also examined the median of the 93 DAC n-tuples for each of these 28 leukemia attractors and found the same results. We also computed the DAC of each attractor against the median of the 1396 apoptosis attractors(Leukemia - Apoptosis), expecting that the DAC for the APOPTOSIS node would be negative. We found the same results as the previous two DAC comparisons in that the DAC of the RONs of the Leukemia associated attractors against the Apoptosis associated attractors was split evenly between positive and negative values.

Conclusions: These results indicate that our clustering method is not clustering phenotypically similar attractors together and thus we cannot use it to associate randomly generated attractors to phenotypes. However, it is not clear if this is due to the output of SFA or the clustering algorithm.

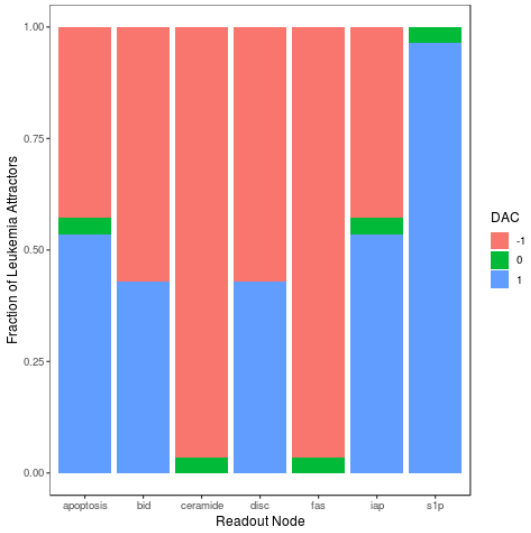


Figure 7. DAC of each RON for the 28 Leukemia attractors that always clustered together with k-means. The values for Ceramide, Fas, and S1P are similar among all the attractors, but the other RONs show varying direction of activity changes.

A Closer Look at the SFA Output.

Perhaps the reason that k-means does not separate Apoptosis and Leukemia associated SFA attractors is because the SFA outputs are not representative of the Apoptosis or Leukemia phenotype. Recall that the 100,000 random initial states lead to 877 attractors in the Boolean model. Out of these 877 Boolean attractors, 133 are Apoptosis associated and 31 are Leukemia associated based on their RONs. In the previous analyses, we only considered states in the basin of the three attractors identified by Zañudo and Albert. To fully understand if we can associate the SFA output to the Apoptosis and Leukemia phenotypes, we wanted to examine the SFA output for all the states in the basin of any Apoptosis or Leukemia associated attractor of the Boolean model. **Figure 8** shows the distribution of the log steady state output estimated by SFA from these initial conditions on the unperturbed network. Clearly, the distribution of estimated activity level of Apoptosis, BID, and IAP does not greatly differ between attractors estimated from initial states in the basin of Leukemia and Apoptosis associated attractors of the Boolean model. Therefore, we cannot expect k-means clustering to distinguish between Apoptosis and Leukemia associated attractors when the SFA outputs for the nodes are similar. It is important to note that these results are very different from the attractors simulated with SFA from a network with fixed nodes (**Figure 4**).

Conclusions: These results indicate that the attractors predicted by SFA applied to the unperturbed network with random initializations may not be associated to the correct phenotype by k-means clustering because the attractor values of RONs of attractors that should represent different phenotypes are similar.

Exploring Other Clustering Options.

To identify the best clustering method for the association of attractors to phenotypes, we explored several dimension reduction techniques on the attractors from the Boolean model. We ran analyses on three separate datasets: 1) The Boolean attractors, 2) the DAC against the Leukemia associated attractor from L1, and 3) the combination of the two. The first approach we took was Principle Component Analysis (PCA). **Figure 9** shows that the 31 Leukemia and the 133 Apoptosis Boolean attractors group separately when graphed with principle component 1 and 2. Figures for PCA on the other two datasets display similar trends.

We also explored both metric and non-metric Multi Dimensional Scaling (MDS) using Manhattan, Canberra, Maximum, Minkowski, Binary, and Hamming distance [9]. This was completed using the R packages MASS and e1071.

MDS uses stress to determine the badness of fit of the dimension reduction. Stress is the normed sum of squares aggregating the representation errors of the model compared to the underlying data [10]. Similarly to PCA, to determine the optimal number of dimensions, we create a scree plot of the stress as the number of dimensions increase and chose the number of dimensions for which further increase in dimensions does not significantly decrease the stress.

Atlantis uses Sammon mapping, a non-linear mapping for non-metric MDS, to plot the attractor landscape by clustering related network states together before and projecting them onto a Cartesian plane [11, 12]. We found that using Sammon mapping on the hamming distance between the Boolean attractors could reduce the data to 3 dimensions (**Figure 10a**). Plotting these three dimensions shows that the Apoptosis attractors cluster separately from the Leukemia attractors (**Figure 10b**).

We have seen that multiple dimension reduction techniques can be applied to the Boolean attractors, and that when plotted using the reduced dimensions, the Leukemia and Apoptosis associated attractors separate. The python module scikit-learn provides examples of applying the k-means algorithm to the reduced dimensions to identify clusters in the data. When applying k-means 2 dimensions from PCA and the 3 dimensions from Sammon mapping, we are able to cluster the Apoptosis associated attractors completely separately from the Leukemia associated attractors, which we were not able to do when using the entire data set for k-means.

It would be ideal to be able to apply a dimension reduction method to the SFA attractors before clustering in the hopes that the Apoptosis associated SFA attractors and the Leukemia associated SFA attractors would then cluster separately as we observed with the Boolean attractors. However, when applying PCA to the "discretized" 100,000 random SFA attractors, the first 3 components only account for 22% of the variance, indicating that it is not appropriate for the data set (**Figure 11**). The same was true for the "discretized" DAC of the SFA attractors and the combination of the steady state and DAC. MDS could not be implemented in R due to the size of the dataset (100,000 attractors). A method for applying Sammon mapping in Python should be explored to determine if SFA attractors can be clustered through MDS.

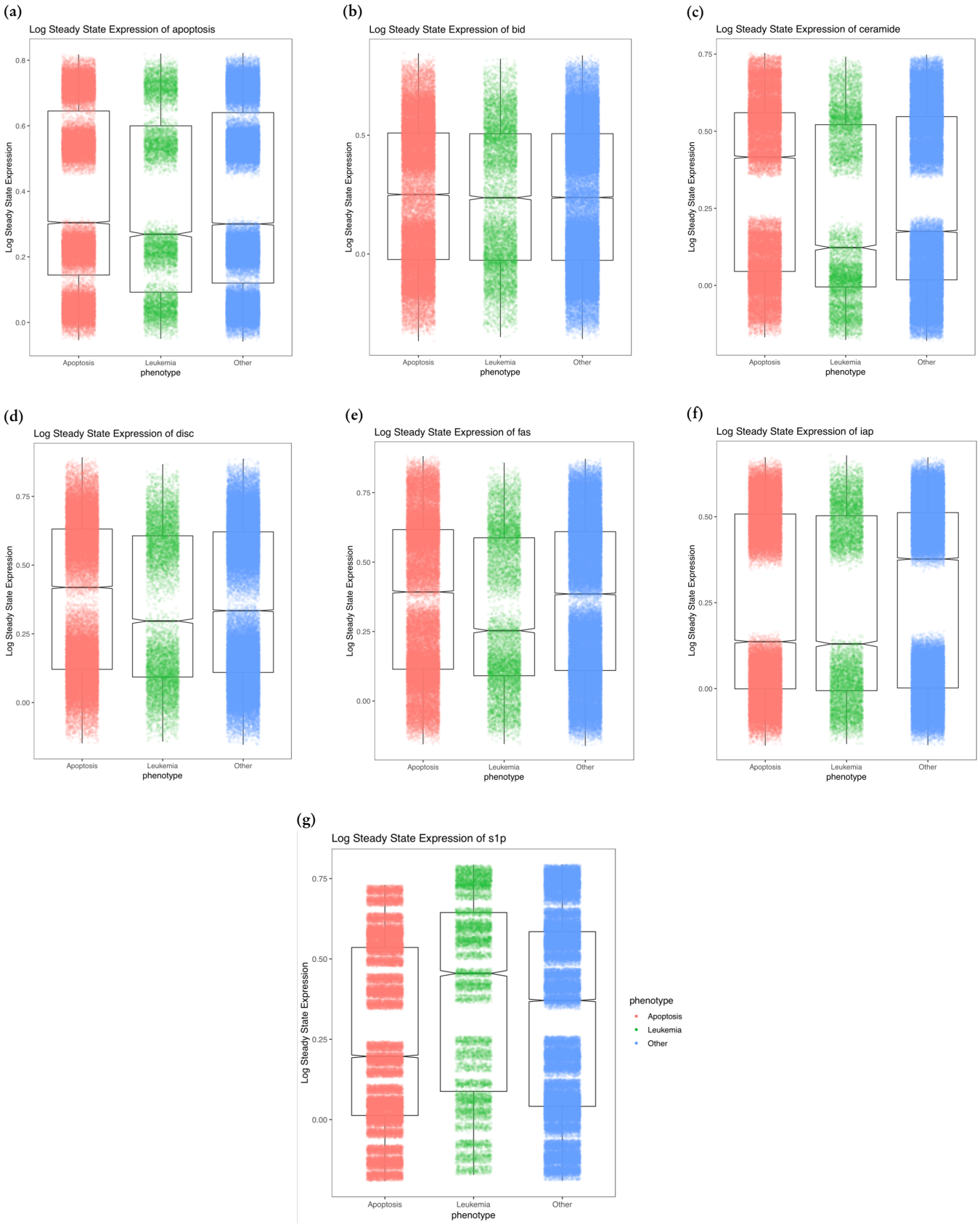
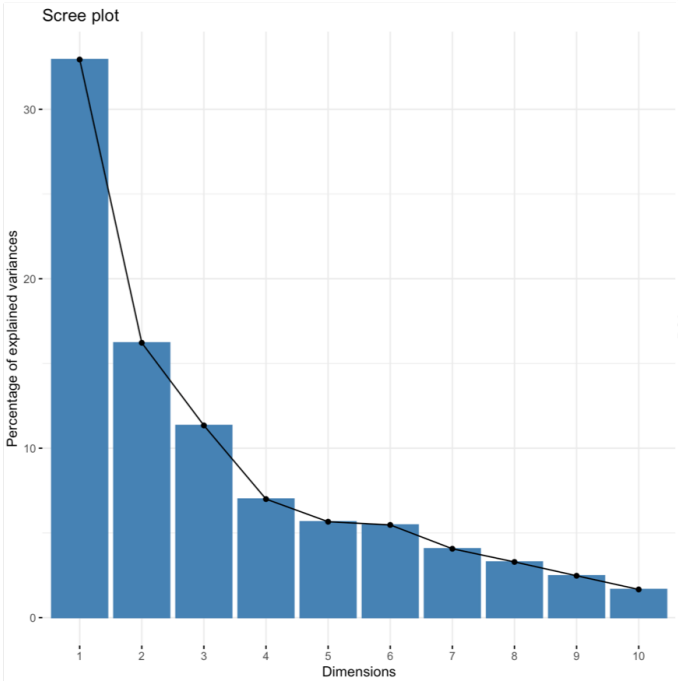


Figure 8. Boxplots of the log steady state value produced by SFA for each the seven RONs split by associated phenotype. Attractors labeled as apoptosis are those from initial states in the basin of Apoptosis attractors in the Boolean model. Likewise, those labeled Leukemia are from initial states in the basin of Leukemia attractors in the Boolean model.

(a) Scree plot for PCA



(b) PCA on Boolean Attractors

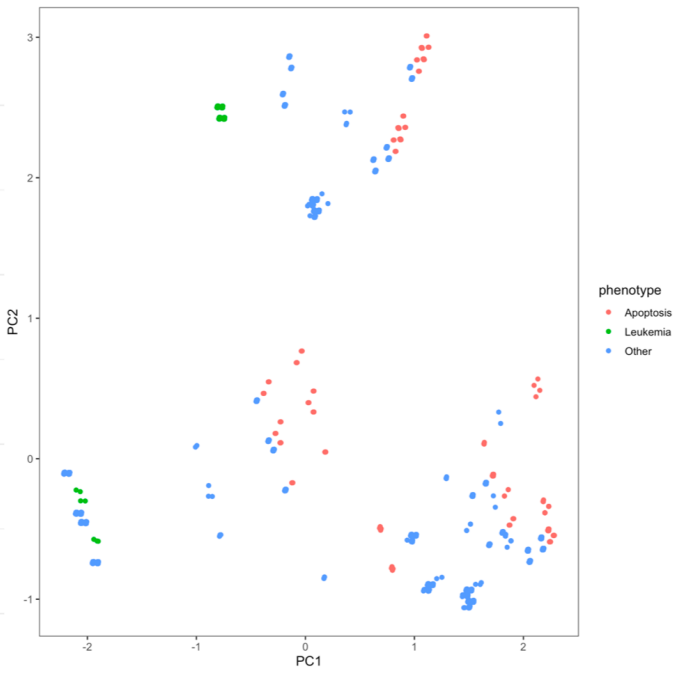
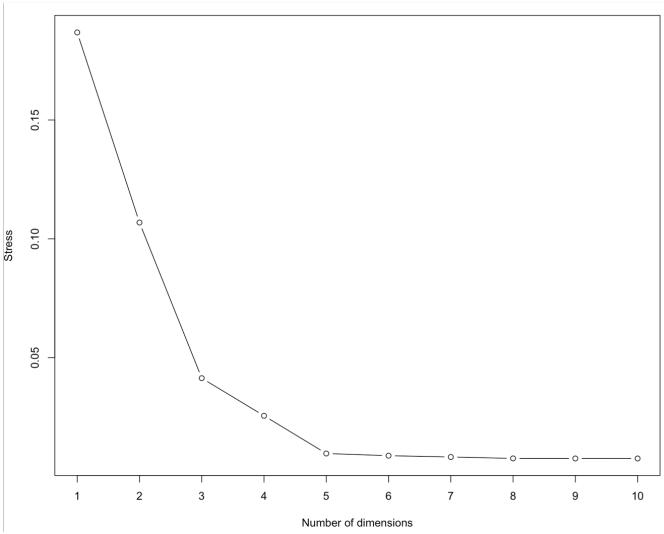


Figure 9. a) The scree plot shows that the first two components capture approximately 50% of the variance in the data. b) Results of PCA applied to the attractors of the Boolean model. Attractors are colored coded based on their associated phenotype determined by the activity of the readout nodes.

(a) Scree plot of stress from Sammon mapping



(b) Sammon mapping on Boolean Attractors with 3 dimensions

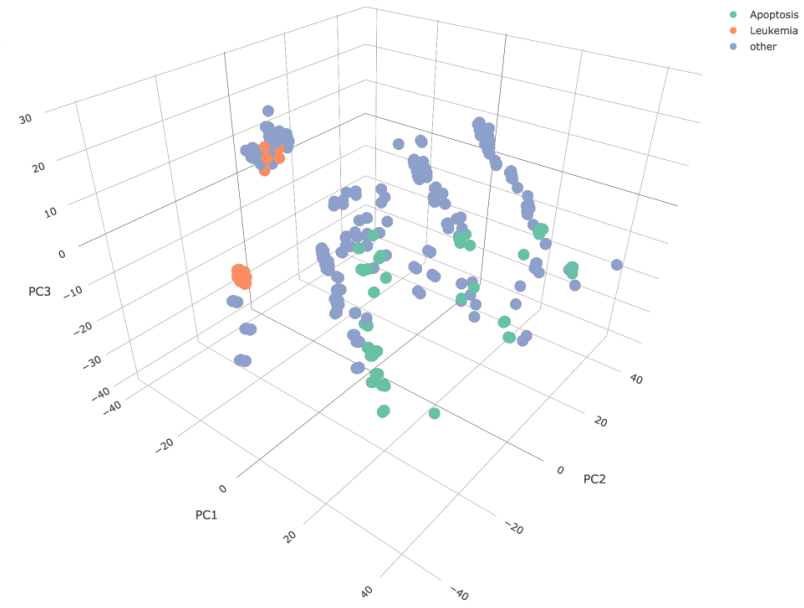


Figure 10. a) Sammon Mapping on the Hamming Distance between Boolean attractors. The stress plot shoes that after the first 3 dimensions, the stress does not significantly decrease as more dimensions are added. b) Results of non-metric MDS sammon mapping on the Boolean attractors.

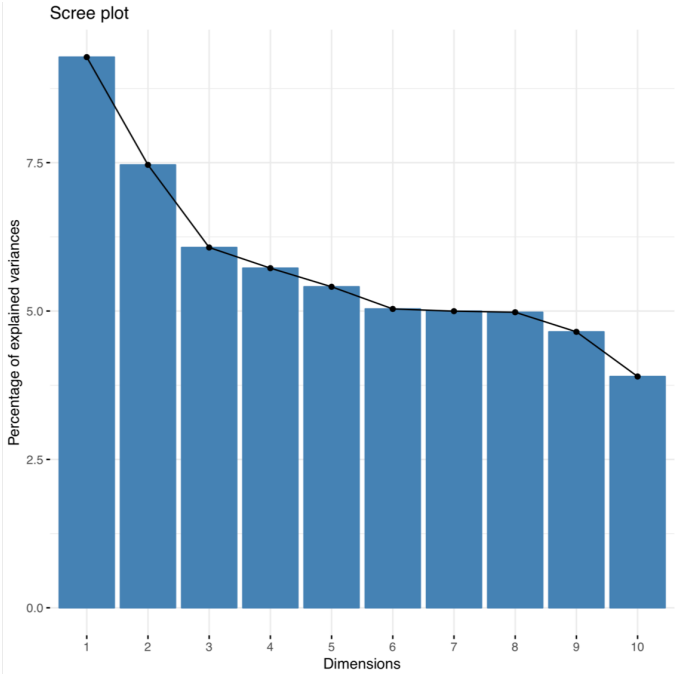


Figure 11. The scree plot for PCA on the SFA attractors show that the first three components do not explain enough of the variance and thus it is not appropriate to apply PCA.

CONCLUSIONS

This exploration has shown that there are cases when we can use SFA as we have been and others when we cannot. **Part 1** shows that when simulating with fixed nodes, SFA can accurately predict the phenotype resulting from these perturbations. Therefore, when we run virtual screenings on our FC node perturbations, we will be able to accurately predict the effect of those perturbations in terms of which nodes are up-regulated and down-regulated. However, the way we currently use SFA is better used to determine if two attractors are different than if they are the same because it cannot predict direction of activity changes of zero. In other words, the way we use SFA is not good at capturing small difference because we discretize the DAC to 1 or -1, regardless of the magnitude of the change. Hence, we need to consider a new interpretation of the output to identify similar attractors.

The analysis from **Part 2** shows that the SFA output does not accurately capture phenotypes when given random initial conditions. In the Boolean network there are network dynamics to steer the trajectory towards an attractor, but since SFA only considers network topology, there is an aspect of the network that is sending SFA predictions away from Leukemia or Apoptosis associated attractors. This is evidenced by the similar distribution of SFA steady state values of RONs between attractors resulting from initial states that laed to Leukemia and Apoptosis associated attractors in the Boolean model (**Figure 8**).

Instead of initializing SFA with 100,000 random conditions, we may need to use biologically relevant values for the source nodes, and 100,000 perturbations of the FVS nodes. Hopefully, this will guide the SFA results towards biologically relevant attractors. Another alternative is to initialize with continuous values as opposed to discrete 0s and 1s. If this does not work, we can at least use SFA to predict the effect of concerted perturbations of the FVS nodes to prioritize perturbations for experimental validation.

REFERENCES

[1] D. Lee and K. H. Cho, "Topological estimation of signal flow in complex signaling networks," *Scientific Reports*, vol. 8, pp. 1–11, 12 2018.

[2] R. Zhang, M. V. Shah, J. Yang, S. B. Nyland, X. Liu, J. K. Yun, R. Albert, and T. P. Loughran, "Network model of survival signaling in large granular lymphocyte leukemia," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp. 16308–16313, 10 2008.

[3] A. Saadatpour, R.-S. Wang, A. Liao, X. Liu, T. P. Loughran, I. Albert, and R. Albert, "Dynamical and Structural Analysis of a T Cell Survival Network Identifies Novel Candidate Therapeutic Targets for Large Granular Lymphocyte Leukemia," *PLoS Computational Biology*, vol. 7, p. e1002267, 11 2011.

[4] J. G. Zañudo and R. Albert, "Cell Fate Reprogramming by Control of Intracellular Network Dynamics," *PLoS Computational Biology*, vol. 11, 4 2015.

[5] C. Müssel, M. Hopfensitz, and H. A.
}Kestler, "BoolNet âĂŞ an R package for generation, reconstruction and
}analysis of Boolean networks," *Bioinformatics*, vol. 26, no. 10, pp. 1378–1380, 2010.

[6] C. A. Gallo, R. L. Cecchini, J. A. Carballido, S. Micheletto, and I. Ponzoni, "Discretization of gene expression data revised," *Briefings in Bioinformatics*, vol. 17, pp. 758–770, 9 2015.

[7] Y. Li, T. Jann, and P. Vera-Licona, "Benchmarking time-series data discretization on inference methods," *Bioinformatics*, vol. 35, pp. 3102–3109, 1 2019.

[8] J. Berthelet and L. Dubrez, "Regulation of Apoptosis by Inhibitors of Apoptosis (IAPs)," *Cells*, vol. 2, pp. 163–187, 3 2013.

[9] J. B. Kruskal, "Nonmetric multidimensional scaling: A numerical method," *Psychometrika*, vol. 29, pp. 115–129, 6 1964.

[10] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, pp. 1–27, 3 1964.

[11] O. S. Shah, M. F. A. Chaudhary, H. A. Awan, F. Fatima, Z. Arshad, B. Amina, M. Ahmed, H. Hameed, M. Furqan, S. Khalid, A. Faisal, and S. U. Chaudhary, "ATLANTIS - Attractor Landscape Analysis Toolbox for Cell Fate Discovery and Reprogramming," *Scientific Reports*, vol. 8, pp. 1–11, 12 2018.

[12] J. W. Sammon, "A Nonlinear Mapping for Data Structure Analysis," *IEEE Transactions on Computers*, vol. C-18, no. 5, pp. 401–409, 1969.