BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 2 2008, pages 293–295
doi:10.1093/bioinformatics/btm571

Systems biology

NET-SYNTHESIS: a software for synthesis, inference and simplification of signal transduction networks

Sema Kachalo[1], Ranran Zhang[2], Eduardo Sontag[3], Réka Albert[4] and Bhaskar DasGupta[5,*]

[1]Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, [2]Penn State Cancer Institute and Integrative Biosciences Graduate Program, Pennsylvania State University, Hershey, PA 17033, [3]Department of Mathematics, Rutgers University, New Brunswick, NJ 08903, [4]Departments of Physics and Biology, Pennsylvania State University, University Park, PA 16802 and [5]Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA

Received on August 21, 2007; revised on November 2, 2007; accepted on November 9, 2007

Advance Access publication November 22, 2007

Associate Editor: Thomas Lengauer

ABSTRACT

Summary: We present a software for combined synthesis, inference and simplification of signal transduction networks. The main idea of our method lies in representing observed indirect causal relationships as network paths and using techniques from combinatorial optimization to find the sparsest graph consistent with all experimental observations. We illustrate the biological usability of our software by applying it to a previously published signal transduction network and by using it to synthesize and simplify a novel network corresponding to activation-induced cell death in large granular lymphocyte leukemia.

Availability: NET-SYNTHESIS is freely downloadable from `http://www.cs.uic.edu/~dasgupta/network-synthesis/`

Contact: dasgupta@cs.uic.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Identification of every reaction and regulatory interaction participating even in a relatively simple function of a single-celled organism requires a concerted and decades-long effort. Consequently, the state of the art understanding of many signaling processes is limited to the knowledge of key mediators and of their positive or negative effects on the whole process. For example, evidence of differential responses to a stimulus in wild-type organisms versus a mutant organism implicates the product of the mutated gene in the signal transduction process. The resulting causal inference relates three components (the signal, the mutated gene and the response) and only in a minority of cases corresponds to a single reaction (namely, when the stimulus is the reactant of the reaction, the mutated gene encodes the enzyme catalyzing the reaction and the studied output is the product of the reaction). We previously introduced (Albert *et al.*, 2007) a method of synthesizing interactions and causal inferences into a parsimonious network by incorporating positive (activating) or negative (inhibitory) causal relationships as signed network paths with known starting and end vertices (nodes) and putative intermediary pseudonodes. Here, we describe an automated version of the method available for use by the community.

*To whom correspondence should be addressed.

2 SOFTWARE OVERVIEW

Our software uses as input a text file whose lines represent causal relationships such as '$A \rightarrow B$' (representing activation), '$A \dashv B$' (representing inhibition), or '$A \rightarrow (B \dashv C)$' (indicating a double causal inference). Relationships that correspond to direct interactions are specified by the label 'Y', e.g. '$A \rightarrow B$ Y'. In addition, the relationship between the enzyme (E) and product (P) of a chemical reaction (i.e. '$E \rightarrow P$') is labeled both 'Y' and 'E' (for enzymatic edge). The entire network synthesis procedure is given in the Supplementary Material; here we briefly describe some key steps. Double causal relationships of the form $A\ x\ (B\ y\ C)$ with $x,y \in \{\rightarrow, \dashv\}$ are represented by adding a new 'pseudo-vertex' P and three new edges, $A\ x\ P$, $B\ a\ P$ and $P\ b\ C$, where a and b are determined by y. Two graph-theoretic procedures, the pseudo-vertex collapse (PVC) and binary transitive reduction (BTR), are used as key steps in the algorithm. Intuitively, the PVC problem is useful for reducing the pseudo-vertex set to the minimal set that maintains the graph consistent with all indirect experimental observations and the BTR problem is useful for determining a sparsest graph consistent with all experimental observations. Although the initial motivation for introducing pseudonodes is to represent the intersection of the two paths corresponding to three-node inferences, PVC can be used in the broader context of network simplification. In many large-scale regulatory networks only a subset of the nodes are of inherent interest, e.g. because they are differentially expressed in different exogenous conditions, and the rest serve as background or mediators. Our software enables users to designate vertices of less interest or confidence as pseudo-vertices and then collapse them, thereby making the network among high-interest/confidence nodes

easier to interpret. To allow gradual simplification, we also provide the choice to collapse degree two pseudonodes only or only collapse one pair of equivalent pseudo-vertices. A detailed manual of the software is available from the software's website. The software should run on any machine with MS Windows (Win32). The source files for a non-graphic version of the program for LINUX/UNIX systems can be obtained by sending an email to the authors.

2.1 Data sources

Large-scale repositories such as Many Microbe Microarrays (`http://m3d.bu.edu/cgi-bin/web/array/index.pl?read=aboutM3D`), NASCArrays (`http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl`) and Gene Expression Omnibus (`http://www.ncbi.nlm.nih.gov/geo/`) contain expression information for thousands of genes under tens to hundreds of experimental conditions. Network inference algorithms applied to gene expression data based on, e.g. mutual information, regression or Bayesian analysis lead to indirect causal relationships among genes. NET-SYNTHESIS can be used to filter redundant inferred relationships by binary transitive reduction. In addition, information about differentially expressed genes responding to a combination of two experimental perturbations, e.g. the presence of a signal in normal versus mutant organisms, can be expressed as double causal inferences. NET-SYNTHESIS can be used to interpret these inferences by pseudo-vertex collapse. Signal transduction pathway repositories such as TRANSPATH (`http://www.gene-regulation.com/pub/databases.html#transpath`) and protein interaction databases such as the Search Tool for the Retrieval of Interacting Proteins (`http://string.embl.de/`) contain up to thousands of interactions, a large number of which are not supported by direct physical evidence. NET-SYNTHESIS can be used to filter redundant information while keeping all direct interactions.

3 RESULTS AND DISCUSSIONS

3.1 Synthesizing a network for T-cell survival and death in large granular lymphocyte leukemia

T-cell large granular lymphocyte leukemia (T-LGL) represents a spectrum of lympho-proliferative diseases in which cytotoxic T lymphocyte activation and elimination are uncoupled (Loughran, 1993). To date, 33 proteins and small molecules related to cytotoxic T lymphocyte activation and activation-induced cell death have been shown to be deregulated in T-LGL and it is known that pro-survival signaling pathways are upregulated and that T-LGL cells are insensitive to Fas-induced apoptosis (Epling-Burnette *et al.*, 2004). However, the interaction/regulatory network among these components remains largely unknown.

We synthesized a cell-survival/cell-death regulation-related signaling network from the TRANSPATH 6.0 database with additional information manually curated from literature search. The 359 vertices of this network represent proteins/protein families and mRNAs participating in pro-survival and Fas-induced apoptosis pathways. The 1295 edges represent regulatory relationships between nodes, including protein interactions, catalytic reactions, transcriptional regulation (for a total of 766 direct interactions) and known indirect causal regulation. No double causal inferences (relationships among three nodes) were available for this network.

Performing BTR with NET-SYNTHESIS reduced the total edge-number to 873. To focus on pathways that involve the 33 known T-LGL deregulated proteins, we designated vertices that correspond to proteins with no evidence of being changed during T-LGL as pseudo-vertices and deleted the label 'Y' for those edges whose both endpoints were pseudo-vertices. Recursively performing 'Reduction (faster)' BTR and 'Collapse degree-2 pseudonodes' of NET-SYNTHESIS until no edge/node could be further removed simplified the network to 267 nodes and 751 edges. Performing comprehensive PVC led to a drastic reduction to 38 vertices and 108 edges. The drawback of this dramatic simplification is that pairs of incoherent edges (two edges with opposite signs) can appear among pairs of nodes. While incoherent paths between pairs of nodes are often seen in biological regulatory networks, interpretation of incoherent edges is difficult without knowledge of the mediators of the two opposite regulatory mechanisms. The number of incoherent edge pairs ranged between 3 (when collapsing degree two pseudo-vertices only) and 19 (for comprehensive PVC). Thus, optimal simplification may require several alternative applications of the various options of PVC algorithms.

3.2 Synthesizing a network for abscisic acid(ABA)-induced stomatal closure

We have performed a comparison of the manually curated network for ABA-induced closure published in Li *et al.* (2006) with the output of NET-SYNTHESIS as reported in Albert *et al.* (2007). The input to NET-SYNTHESIS is a list of 140 interactions and causal inferences in ABA-induced closure published in Table S1 and Text S1 in Li *et al.* (2006). The complete list of causal relationships is given in Table 1 in the Supplementary Material. A detailed comparison of the two networks is available in Albert *et al.* (2007), here we briefly summarize the overall comparison of the two networks. The network of Li *et al.* (2006) has 54 vertices and 92 edges; our network has 57 vertices (3 extra pseudo-vertices) but 84 edges. The two networks have 71 common edges and identical strongly connected components. All the paths present in the (Li *et al.*, 2006) reconstruction are present in our network as well. Thus, the two networks are highly similar and their divergence on a few edges is due not to algorithmic deficiencies but to human decisions. Finally, the entire network synthesis process was done within a few seconds by our software. A picture of our network is available as Figure 1 in the Supplementary Material.

4 CONCLUSION

The applications of NET-SYNTHESIS enable us to conclude that it can serve as a very important first step in formalizing the logical substrate of an inferred signal transduction network. We foresee its optimal application in conjunction with human expertise, as part of an interactive and iterative process. The NET-SYNTHESIS users would give the experimentally known

information as input, then use the output network to augment the input information with additional facts or hypotheses, allowing them to simultaneously synthesize their knowledge and formalize their hypotheses regarding a signal transduction network.

ACKNOWLEDGEMENTS

This project was partially supported by NSF grants IIS-0346973 (to S.K.), EIA-0205116 and DMS-050455 (to E.D.S.), MCB-0618402 and CCF-0643529 (to R.A.), IIS-0346973, IIS-0612044 and DBI-0543365 (to B.D.) and USDA grant 2006-35100-17254 (to R.A.).

Conflict of Interest: none declared.

REFERENCES

Albert,R. *et al.* (2007) A novel method for signal transduction network inference from indirect experimental evidence. *J. Comp. Biol.*, **14**, 927–949.

Epling-Burnette,P.K. *et al.* (2004) ERK couples chronic survival of NK cells to constitutively activated Ras in lymphoproliferative disease of granular lymphocytes. *Oncogene*, **23**, 9220–9229.

Li,S. *et al.* (2006) Predicting essential components of signal transduction networks: a dynamic model of guard cell abscisic acid signaling. *PLoS Biol.*, **4**, e312, doi:10.1371/journal.pbio.0040312.

Loughran,T.P., Jr (1993) Clonal diseases of large granular lymphocytes. *Blood*, **82**, 1–14.

interaction	critical	enzymatic	interaction	critical	enzymatic
ABA → SphK	No	No	ABA → OST1	No	No
ABA → CaIM	No	No	ABA → InsP6	No	No
ABA → Ca2+c	No	No	ABA → NO	No	No
ABA → InsP3	No	No	ABA → AnionEM	No	No
ABA ⊣ PEPC	No	No	ABA ⊣ Malate	No	No
ABA ⊣ HATPase	No	No	ABA ⊣ RAC1	No	No
ABA → PLD	No	No	ABA → ROS	No	No
Ca2+c ⊣ CaIM	No	No	Ca2+c → KEV	No	No
Ca2+c → AnionEM	No	No	InsP6 → Ca2+c	No	No
InsP6 → CIS	No	No	ROS → CaIM	No	No
ROS → Closure	No	No	ROS ⊣ ABI1	No	No
ROS ⊣ KOUT	No	No	pHc → KOUT	No	No
pHc → ABI1	No	No	pHc → ROS	No	No
pHc → HATPase	No	No	PA ⊣ ABI1	No	No
PA → Closure	No	No	PA → ROS	No	No
NO → Closure	No	No	NO → AnionEM	No	No
NO ⊣ KOUT	No	No	RAC1 ⊣ Actin	No	No
RAC1 ⊣ Closure	No	No	ABH1 ⊣ AnionEM	No	No
AnionEM ⊣ Malate	No	No	ERA1 → ROP10	No	No
Depolarization ⊣ Ca2+c	No	No	GPA1 → PLD	Yes	No
Sph → S1P	Yes	No	InsPK → InsP6	Yes	Yes
PLC → DAG	Yes	Yes	PIP2 → DAG	Yes	No
PLC → InsP3	Yes	Yes	PIP2 → InsP3	Yes	No
GC → cGMP	Yes	Yes	GTP → cGMP	Yes	No
ADPRc → cADPR	Yes	Yes	NAD → cADPR	Yes	No
NADPH → NO	Yes	No	Nitrite → NO	Yes	No
Arg → NO	Yes	No	NOS → NO	Yes	Yes
NIA12 → NO	Yes	Yes	NADPH → ROS	Yes	No
Atrboh → ROS	Yes	Yes	Ca2+ATPase ⊣ Ca2+c	Yes	No
Ca2+c → Ca2+ATPase	Yes	No	HATPase ⊣ Depolarization	Yes	No
KOUT ⊣ Depolarization	Yes	No	KAP ⊣ Depolarization	Yes	No
AnionEM → Depolarization	Yes	No	Ca2+c → Depolarization	Yes	No
KEV → Depolarization	Yes	No	RCN1 → NIA12	No	No
CIS → Ca2+c	Yes	No	CaIM → Ca2+c	Yes	No
Malate ⊣ Closure	Yes	No	GCR1 ⊣ GPA1	Yes	No
ABA → RCN1	No	No	AnionEM → Closure	Yes	No
KAP → Closure	Yes	No	KOUT → Closure	Yes	No
ERA1 ⊣ CaIM	No	No	ABH1 ⊣ CaIM	No	No
cGMP → CIS	No	No	cADPR → CIS	No	No
InsP3 → CIS	No	No	Ca2+c → NOS	No	No
ROS → (ABA → Closure)	-	-	AnionEM → (ABA → Closure)	-	-
PLC → (ABA → Closure)	-	-	SphK → (ABA → Closure)	-	-
SphK → (ABA → AnionEM)	-	-	SphK → (ABA → S1P)	-	-
S1P → (ABA → Closure)	-	-	GPA1 → (S1P → AnionEM)	-	-
GPA1 → (ABA → ROS)	-	-	GCR1 ⊣ (ABA → Closure)	-	-
PLC → (ABA → Ca2+c)	-	-	cADPR → (ABA → Ca2+c)	-	-
NOS → (ABA → Closure)	-	-	NO → (ABA → Closure)	-	-
NO → (ABA → Closure)	-	-	NO → (ABA → AnionEM)	-	-
Ca2+c → (NO → AnionEM)	-	-	NO → (Ca2+c → CIS)	-	-
ADPRc → (NO → Ca2+c)	-	-	GC → (NO → Ca2+c)	-	-
KOUT → (ABA → Closure)	-	-	GPA1 → (ABA → AnionEM)	-	-
pHc → (ABA → Closure)	-	-	ERA1 ⊣ (ABA → AnionEM)	-	-
ERA1 ⊣ (ABA → Closure)	-	-	ERA1 ⊣ (Depolarization → KOUT)	-	-
Atrboh → (ABA → Closure)	-	-	Atrboh → (ABA → ROS)	-	-
Atrboh → (ABA → Ca2+c)	-	-	Atrboh → (ABA → CaIM)	-	-
ROS → (ABA → CaIM)	-	-	NADPH → (ABA → CaIM)	-	-
NAD → (ABA → CaIM)	-	-	ERA1 ⊣ (ABA → CaIM)	-	-
ERA1 ⊣ (ABA → Closure)	-	-	RCN1 → (ABA → Closure)	-	-
RCN1 → (ABA → AnionEM)	-	-	RCN1 → (ABA → Ca2+c)	-	-
OST1 → (ABA → Closure)	-	-	OST1 → (ABA → ROS)	-	-
PLC → (ABA → Closure)	-	-	Ca2+c → (ABA → Closure)	-	-
AnionEM → (ABA → Closure)	-	-	PLD → (PC → PA)	-	-
PLD → (ABA → Closure)	-	-	PLC → (ABA → Closure)	-	-
ABA → (PLD → PA)	-	-	ABA → (PLD → PA)	-	-

Table 1. Regulatory interactions between ABA signal transduction pathway components (Li et al., 2006).

Table 1 (continued)

interaction	critical	enzymatic	interaction	critical	enzymatic
ROP2 → (PA → ROS)	-	-	Actin → (ABA → Closure)	-	-
Ca2+c → (ABA → Actin)	-	-	RAC1 ⊣ (ABA → Closure)	-	-
ROP10 ⊣ (ABA → Closure)	-	-	ROS → (ABA → Closure)	-	-
GCR1 ⊣ (ABA → Closure)	-	-	GCR1 ⊣ (S1P → Closure)	-	-
cADPR → (Ca2+c → CIS)	-	-	AnionEM → (ABA → Closure)	-	-
CaIM → (ABA → KOUT)	-	-	cADPR → (ABA → KOUT)	-	-
PLC → (ABA → KOUT)	-	-	ROS → (ABA → CaIM)	-	-
Ca2+c ⊣ (Depolarization → KAP)	-	-	pHc ⊣ (Depolarization → KAP)	-	-
ABH1 ⊣ (ABA → Closure)	-	-	ABH1 ⊣ (ABA → Ca2+c)	-	-
ROS → (ABA ⊣ HATPase)	-	-	ABI1 ⊣ (ABA → AnionEM)	-	-
ABI1 ⊣ (ABA → ROS)	-	-	ABI1 ⊣ (ABA → Ca2+c)	-	-
AtPP2C ⊣ (ABA → Closure)	-	-	Ca2+c → (PLC → InsP3)	-	-
GPA1 → AGB1	No	No	AGB1 → GPA1	No	No
AtPP2C ⊣ Closure	No	No	NO → ADPRc	No	No
Ca2+c → HATPase	No	No	ABI1 ⊣ Atrboh	No	No
NO → GC	No	No	ABA → pHc	No	No
PA → ROP2	No	No	PEPC → Malate	Yes	Yes
ABI1 ⊣ (ABA → ROS)	-	-	ABA → PLC	No	No
Depolarization → KOUT	Yes	No	Depolarization → KAP	Yes	No
Depolarization ⊣ CaIM	Yes	No	ABI1 → (ABA ⊣ RAC1)	-	-
InsPK → (ABA → AnionEM)	-	-	InsPK → (ABA → InsP6)	-	-
S1P → GPA1	No	No			

Table 1. Regulatory interactions between ABA signal transduction pathway components (Li et al., 2006).

Supplementary Information: Description of the Network Synthesis Procedure

Here we sketch the framework of the network synthesis procedure employed in NET-SYNTHESIS. A complete description can be found in [Albert *et al.*, 2007].

The procedure applies to directed graphs $G = (V, E)$ with an edge labeling function $w : E \mapsto \{0, 1\}$. Biologically, edge labels 0 and 1 in edges $u \xrightarrow{0} v$ and $u \xrightarrow{1} v$ correspond to "u promotes v" and "u inhibits v", respectively.

The *parity* (sign) of a path P from vertex u to vertex v is $\sum_{e \in P} w(e) \pmod 2$. A path of parity 0 is called a path of *even* parity, or positive sign. A path of parity 1 is called a path of *odd* parity, or negative sign. The notation $u \overset{x}{\Rightarrow} v$ denotes a path from u to v of parity $x \in \{0, 1\}$.

For a subset of edges $E' \subseteq E$, reachable(E') is the set of all ordered triples (u, v, x) such that $u \overset{x}{\Rightarrow} v$ is a path of the subgraph (V, E'). The binary transitive reduction (BTR) problem is defined as follows:

> **Instance:** A directed graph $G = (V, E)$ with an edge labeling function $w : E \mapsto \{0, 1\}$ and a set of critical edges $E_{\text{critical}} \subseteq E$.
>
> **Valid Solutions:** A subgraph $G' = (V, E')$ where $E_{\text{critical}} \subseteq E' \subseteq E$ and reachable(E') =reachable(E).
>
> **Objective:** *Minimize* $|E'|$.

Intuitively, the BTR problem is useful for determining the sparsest graph consistent with a set of experimental observations. The set of "critical edges" represent edges which are known to be direct interactions with concrete evidence.

The pseudo-vertex collapse (PVC) problem is defined as follows:

> **Instance:** A directed graph $G = (V, E)$ with an edge labeling function $w : E \mapsto \{0, 1\}$ and a subset $V' \subset V$ of vertices called pseudo-vertices. The vertices in $V \setminus V'$ are called "real" vertices.
>
> **Definition:**
>
> - For any vertex v, let $\text{in}(v) = \{(u, x) \mid u \overset{x}{\Rightarrow} v, x \in \{0, 1\}\} \setminus \{v\}$ and let $\text{out}(v) = \{(u, x) \mid v \overset{x}{\Rightarrow} u, x \in \{0, 1\}\} \setminus \{v\}$.
> - Collapsing two vertices u and v is permissible provided both are not "real" vertices and in(u) =in(v) and out(u) =out(v).
> - If permissible, the collapse of two vertices u and v creates a new vertex w, makes every incoming (resp. outgoing) edges to (resp. from) either u or v an incoming (resp. outgoing) edge from w, removes any parallel edge that may result from the collapse operation and also removes both vertices u and v.
>
> **Valid Solutions:** A graph $G'' = (V'', E'')$ obtained from G by a sequence of permissible collapse operations.
>
> **Objective:** *Minimize* $|V''|$.

Intuitively, the PVC problem is useful for reducing the pseudo-vertex set to the the minimal set that maintains the graph consistent with all indirect experimental observations. As in the case of the BTR problem, our goal is to minimize false positive (spurious) inferences of additional components in the network.

The main network synthesis steps employed in NET-SYNTHESIS are the following:

1. Incorporate single causal inferences and biochemical interactions as labeled edges, noting the critical edges corresponding to direct interactions.
2. Perform a binary transitive reduction to eliminate spurious inferred edges (*i.e.*, edges that can be explained by paths of the same label).
3. Incorporate double causal relationships $A \stackrel{x}{\rightarrow} (B \stackrel{y}{\rightarrow} C)$ by (i) adding a new edge $A \stackrel{x}{\rightarrow} B$ if $B \stackrel{y}{\rightarrow} C$ is an existing critical edge, (ii) doing nothing if existing paths in the network already explain the relationship, or (iii) adding a new pseudo-vertex P and three new edges $A \stackrel{x}{\rightarrow} P$, $B \stackrel{a}{\rightarrow} P$ and $P \stackrel{b}{\rightarrow} C$. To correctly incorporate the parity of the $A \stackrel{x+y \pmod 2}{\longrightarrow} C$ relationship, positive $B \stackrel{y}{\rightarrow} C$ paths will be broken into two positive edges, while negative paths will be broken into a positive edge ($a = 0$) and a negative edge ($b = 1$), summarized in a concise way by the equation $b = a + b = y \pmod 2$.
4. Perform pseudo-vertex collapse to reduce unnecessary redundancy in the resulting graph.
5. Perform a second round of binary transitive reduction to eliminate any redundant edges created by pseudo-vertex collapse.

Intuitively speaking, the approach is to first expand the network by the addition of the pseudo-vertices at the intersection of the two paths corresponding to double (three-node) causal inferences, then to use the additional information available in the network to collapse these pseudo-vertices, *i.e.*, to identify them with real vertices or with each other.

NET-SYNTHESIS offers two additional procedures of interest in certain biological networks. If chosen by the user, indirect regulation of a product of an enzymatic reaction will be interpreted as regulation of the enzyme, implemented as the action "Enzymatic edges". The user can also designate any node as a pseudo-node by prepending a * to the name of the node, and then perform PVC either on the whole network or on degree-two pseudo-nodes only. An example of a set of input interactions for a network synthesis approach is given in the file `http://www.cs.uic.edu/~dasgupta/network-synthesis/sample-input-file.txt` on the NET-SYNTHESIS webpage. This file also provides a suggested sequence of actions for this network.

References

[Albert *et al.*, 2007] R. Albert, B. DasGupta, R. Dondi, et al. (2007). A Novel Method for Signal Transduction Network Inference from Indirect Experimental Evidence, Journal of Computational Biology, 14 (7), 927-949.

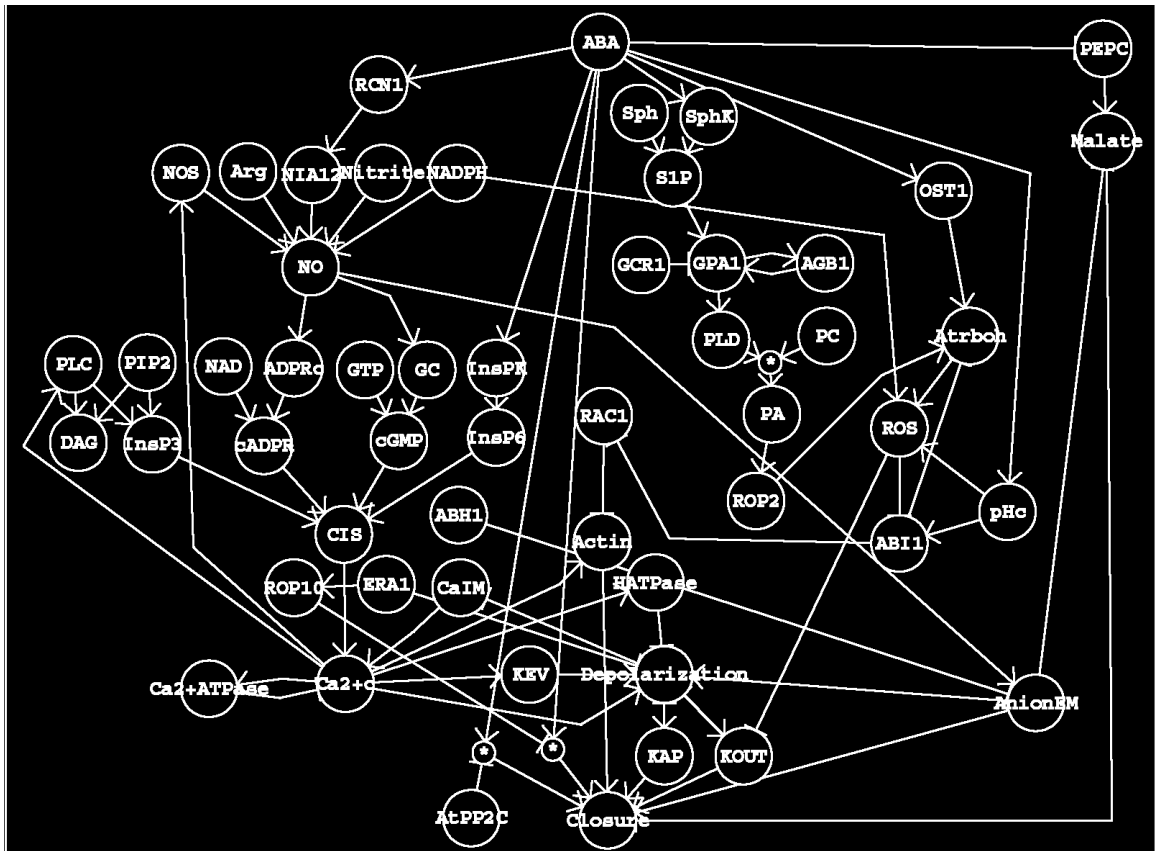

Figure 1: The guard cell signal transduction network for ABA-induced stomatal closure produced by NET-SYNTHESIS.