Constructing a Knowledge Base for Gene Regulatory Dynamics by Formal Concept Analysis Methods

Johannes Wollbold[1 2] Reinhard Guthke[2] Bernhard Ganter[1]

[1]University of Technology, Institute of Algebra, Dresden, Germany
http://www.math.tu-dresden.de/alg/algebra.html
jwollbold@gmx.de

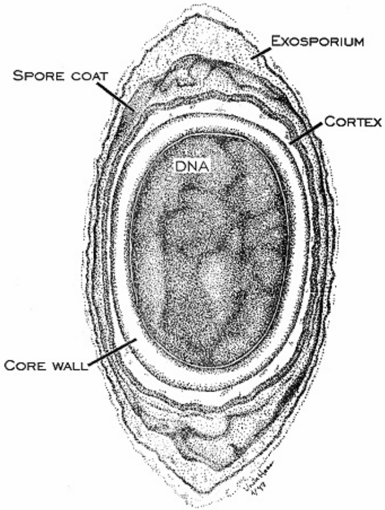
[2]Leibniz Institute for Natural Product Research and Infection Biology - Hans-Knöll-Institute (HKI) Jena, Germany

Algebraic Biology 2008, 1st of August, Hagenberg, AT

Outline

1. The Example: Sporulation in Bacillus subtilis
2. Introduction of 4 Formal Contexts
3. Attribute Exploration of the Transitive Context Representing a Complete Simulation
4. Outlook

Bacillus subtilis



- Gram positive soil bacterium.
- Produces single endospores under environmental stress, which can survive ultraviolet or gamma radiation, acid, hours of boiling, or starvation.
- Switch between two completely different genetic programs.
- Model organism.

Boolean Network for the Sporulation Decision [Ste07]

$$
\begin{aligned}
\text{AbrB} &= \text{SigA}\ \overline{\text{AbrB}}\ \overline{\text{Spo0AP}} \\
\text{SigF} &= (\text{SigH}\ \text{Spo0AP}\ \overline{\text{SinR}}) + (\text{SigH}\ \text{Spo0AP}\ \text{SinI}) \\
\text{KinA} &= \text{SigH}\ \overline{\text{Spo0AP}} \\
\text{Spo0A} &= (\text{SigH}\ \overline{\text{Spo0AP}}) + (\text{SigA}\ \overline{\text{Spo0AP}}) \\
\overline{\text{Spo0A}} &= (\overline{\text{SigA}}\ \text{SinR}\ \overline{\text{SinI}}) + (\overline{\text{SigH}}\ \overline{\text{SigA}}\ ) + \text{Spo0AP} \\
\text{Spo0AP} &= \text{Signal}\ \text{Spo0A}\ \overline{\text{Spo0E}}\ \text{KinA} \\
\text{Spo0E} &= \text{SigA}\ \overline{\text{AbrB}} \\
\text{SigH} &= \text{SigA}\ \overline{\text{AbrB}} \\
\text{Hpr} &= \text{SigA}\ \text{AbrB}\ \overline{\text{Spo0AP}} \\
\text{SinR} &= (\text{SigA}\ \overline{\text{AbrB}}\ \overline{\text{Hpr}}\ \overline{\text{SinR}}\ \overline{\text{SinI}}\ \text{Spo0AP}) + \\
&\quad\ (\text{SigA}\ \overline{\text{AbrB}}\ \overline{\text{Hpr}}\ \text{SinR}\ \text{SinI}\ \text{Spo0AP}) \\
\text{SinI} &= \text{SinR} \\
\text{SigA} &= \text{TRUE (input to the model)} \\
\text{Signal} &= \text{TRUE or FALSE (constant, depending on the input state)}
\end{aligned}
$$

Outline

Formal Contexts and Concepts

Definition

A **formal context** (G, M, I) defines a relation $I \subseteq G \times M$ between objects G and attributes M. For $A \subseteq G$ and $B \subseteq M$, derivation operators $'$ are defined by:

$$A' := \{m \in M |\, gIm \text{ for all } g \in A\}$$

$$B' := \{g \in G |\, gIm \text{ for all } m \in B\}$$

Definition

A **formal concept** of the context (G, M, I) is a pair (A, B) with $A \subseteq G$, $B \subseteq M$, $A' = B$ and $B' = A$. A is the **extent**, B the **intent** of the concept (A, B).

$''$ defines closure operators on G and M.
(A'', A') and (B', B'') are concepts, for all $A \in G$, $B \in M$.

A State Context Representing a Simulation without Stress

Given: sets E (entities) and F (fluents),
states $\varphi \in F_1 \times ... \times F_n,\ n = |E|,\ F_1 = ... = F_n = F$.

Definition

A **state context** is a formal context (G, M, I) with
$G \subseteq F^E := \{\varphi : E \to F\}$ and $M \subseteq E \times F$; its relation I is given as
$\varphi\, I\, (e, f) :\Leftrightarrow \varphi(e) = f$, for all $\varphi \in G, e \in E$ and $f \in F$.

State	KinA	Spo0A	Spo0AP	AbrB	Spo0E	SigH	Hpr	...
φ_0	-	+	-	-	-	-	+	
φ_1	-	+	-	+	+	+	-	
φ_2	+	+	-	-	-	-	+	

A state context in the form of a many-valued context.

A State Context Representing a Simulation without Stress

Given: sets E (entities) and F (fluents),
states $\varphi \in F_1 \times \ldots \times F_n,\ n = |E|,\ F_1 = \ldots = F_n = F$.

Definition

A **state context** is a formal context (G, M, I) with
$G \subseteq F^E := \{\varphi : E \to F\}$ and $M \subseteq E \times F$; its relation I is given as
$\varphi\, I\, (e, f) :\Leftrightarrow \varphi(e) = f$, for all $\varphi \in G, e \in E$ and $f \in F$.

State	KinA.off	KinA.on	Spo0A.off	Spo0A.on	Spo0AP.off	Spo0AP.on	AbrB.off	AbrB.on	Spo0E.off	Spo0E.on	SigH.off	SigH.on	Hpr.off	Hpr.on	...
φ_0	x			x	x		x		x		x			x	
φ_1	x			x	x			x		x		x	x		
φ_2		x		x	x		x		x		x			x	

The state context scaled to a one-valued context (corresponding to Def.).

A State Context Representing a Simulation without Stress

Given: sets E (entities) and F (fluents),
states $\varphi \in F_1 \times \ldots \times F_n$, $n = |E|$, $F_1 = \ldots = F_n = F$.

Definition

A **state context** is a formal context (G, M, I) with
$G \subseteq F^E := \{\varphi : E \rightarrow F\}$ and $M \subseteq E \times F$; its relation I is given as
$\varphi \, I \, (e, f) :\Leftrightarrow \varphi(e) = f$, for all $\varphi \in G, e \in E$ and $f \in F$.

State	KinA.off	KinA.on	Spo0A.off	Spo0A.on	Spo0AP.off	AbrB.off	Spo0E.off	SigH.off	Hpr.on	Spo0AP.on	AbrB.on	Spo0E.on	SigH.on	Hpr.off	...
φ_0	x			x	x	x	x	x	x						
φ_2		x		x	x	x	x	x	x						
φ_1	x			x	x						x	x	x	x	

Concept = maximal rectangle, after permutation of rows and columns.

Transition Context

Definition

Given a state context (G, M, I) and a relation $R \subseteq G \times G$, a **transition context** $\mathbb{K}$ is the context $(R, M \times \{in, out\}, \tilde{I})$ with the property

$$(\varphi_1, \varphi_2)\tilde{I}(e, f, i) :\Leftrightarrow \begin{cases} \varphi_1(e) = f & \text{for } i = in \\ \varphi_2(e) = f & \text{for } i = out. \end{cases}$$

Transition	**KinA.in.off**	**KinA.in.on**	**AbrB.in.off**	**AbrB.in.on**	**Spo0E.in.off**	**Spo0E.in.on**	**SigH.in.off**	**SigH.in.on**	**Hpr.in.off**	**Hpr.in.on**	...	**KinA.out.off**	**KinA.out.on**	**AbrB.out.off**	**AbrB.out.on**	**Sp0E.out.off**	**Sp0E.out.on**	**SigH.out.off**	**SigH.out.on**	**Hpr.out.off**	**Hpr.in.on**	...
(φ_0, φ_1)	x		x		x		x			x		x			x		x		x	x		
(φ_1, φ_2)	x			x		x		x	x				x	x		x		x			x	
(φ_2, φ_1)		x	x		x		x			x		x			x		x		x	x		

Transitive context $\mathbb{K}_t$: transitively closed relation $t(R) := \bigcup_{n \in \mathbb{N} \setminus \{0\}} R^n$.

The Concept Lattice of the Transitive Context

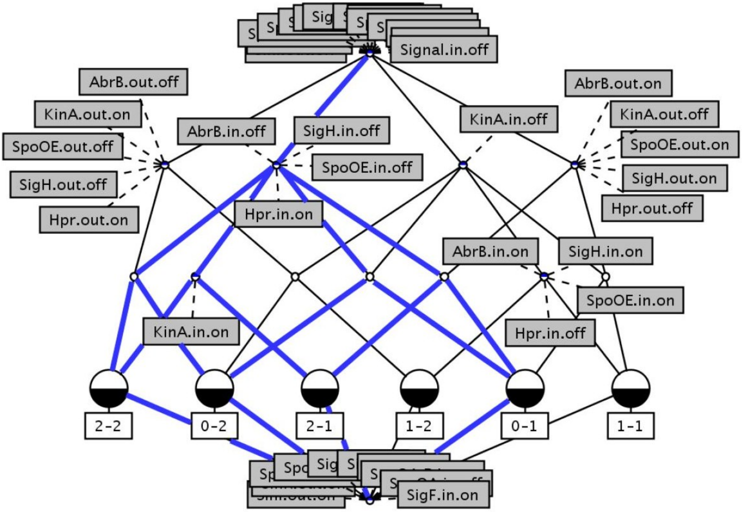


Figure: *Signal*: starvation; *AbrB, Hpr, SigA, SigF, SigH, SinR, Spo0A* (phosporylated form *Spo0AP*): transcription factors; *KinA*: kinase; *Spo0E*: phosphatase; *SinI* inactivates *SinR* by binding to it.

Boolean Network for the Sporulation Decision [Ste07]

$$
\begin{aligned}
\mathrm{AbrB} &= \mathrm{SigA}\ \overline{\mathrm{AbrB}}\ \overline{\mathrm{Spo0AP}} \\
\mathrm{SigF} &= (\mathrm{SigH}\ \mathrm{Spo0AP}\ \overline{\mathrm{SinR}}) + (\mathrm{SigH}\ \mathrm{Spo0AP}\ \mathrm{SinI}) \\
\mathrm{KinA} &= \mathrm{SigH}\ \overline{\mathrm{Spo0AP}} \\
\mathrm{Spo0A} &= (\mathrm{SigH}\ \overline{\mathrm{Spo0AP}}) + (\mathrm{SigA}\ \overline{\mathrm{Spo0AP}}) \\
\overline{\mathrm{Spo0A}} &= (\overline{\mathrm{SigA}}\ \mathrm{SinR}\ \overline{\mathrm{SinI}}) + (\overline{\mathrm{SigH}}\ \overline{\mathrm{SigA}}\ ) + \mathrm{Spo0AP} \\
\mathrm{Spo0AP} &= \mathrm{Signal}\ \mathrm{Spo0A}\ \overline{\mathrm{Spo0E}}\ \mathrm{KinA} \\
\mathrm{Spo0E} &= \mathrm{SigA}\ \overline{\mathrm{AbrB}} \\
\mathrm{SigH} &= \mathrm{SigA}\ \overline{\mathrm{AbrB}} \\
\mathrm{Hpr} &= \mathrm{SigA}\ \mathrm{AbrB}\ \overline{\mathrm{Spo0AP}} \\
\mathrm{SinR} &= (\mathrm{SigA}\ \overline{\mathrm{AbrB}}\ \overline{\mathrm{Hpr}}\ \overline{\mathrm{SinR}}\ \overline{\mathrm{SinI}}\ \mathrm{Spo0AP}) + \\
&\quad (\mathrm{SigA}\ \overline{\mathrm{AbrB}}\ \overline{\mathrm{Hpr}}\ \mathrm{SinR}\ \mathrm{SinI}\ \mathrm{Spo0AP}) \\
\mathrm{SinI} &= \mathrm{SinR} \\
\mathrm{SigA} &= \text{TRUE (input to the model)} \\
\mathrm{Signal} &= \text{TRUE or FALSE (constant, depending on the input state)}
\end{aligned}
$$

Extended State Context

Supplementary temporal attributes always(m), never(m), eventually(m)...

State	KinA	Spo0A	Spo0AP	AbrB	Spo0E	SigH	Hpr	ev(KinA)	alw(KinA)	nev(Spo0AP)	ev(AbrB)	alw(AbrB)	ev(Hpr)	...
φ_0	-	+	-	-	-	-	+	x		x	x		x	
φ_1	-	+	-	+	+	+	-	x		x	x		x	
φ_2	+	+	-	-	-	-	+	x		x	x		x	

Outline

The Attribute Exploration Algorithm

- Interactive algorithm.
- Generates implications $A \to B$ between attribute sets of a given formal context.
- An expert or a computer program decides about the general validity of the rule, e.g. by comparison of a simulation with observed time series.
- No $\curvearrowright$ counterexample = new object.
- Yes $\curvearrowright$ add $A \to B$ to the stem base of the context.
- Sound, complete, non redundant - each valid implication is derivable by the Armstrong rules in linear time ($W, X, Y, Z \subseteq M$):

$$\frac{\top}{X \to X} \qquad \frac{X \to Y}{X \cup Z \to Y} \qquad \frac{X \to Y,\ Y \cup Z \to W}{X \cup Z \to W}$$

- Mathematical foundation: Pseudo-closed sets, Theorem of Duquenne-Guigues [GW99, Theorem 8]

Some Interesting Implications of the Stem Base

- 4224 transitions from all possible $2^{12} = 4096$ initial states.
- 11.700 transitions in the transitive context.
- Stem base: 524 implications with support > 0 (computing time < 12h).
- $11.023.494 \approx 2^{24}$ concepts in the lattice.

SigH.out.off → AbrB.out.off, SpoOE.out.off, SinR.out.off, SinI.out.off

All these genes are coregulated: $\overline{\textit{gene.out}} = \overline{\text{SigA.in}} + \text{AbrB.in}$ (+ ...).

SigF.out.on → KinA.out.off, Spo0A.out.off, Hpr.out.off, AbrB.out.off

Spo0AP is reported to activate the production of SigF but also to repress its own expression (mutual exclusion). [De 04]

Implications Specific for the Transitive Context

$< 4500 >$ Spo0AP.in.on, KinA.out.off $\longrightarrow$ Hpr.out.off
$< 4212 >$ SigH.in.on. KinA.out.off $\longrightarrow$ Hpr.out.off
$< 3972 >$ AbrB.in.off, KinA.out.off $\longrightarrow$ Hpr.out.off

$\overline{\text{Hpr}}$ and $\overline{\text{KinA}}$ are determined by different Boolean functions, but they are coregulated in all states emerging from any input state characterized by the single attributes Spo0AP.on, SigH.on or AbrB.on.

$< 3904 >$ AbrB.out.on
$\longrightarrow$ SigA.in.on, SigA.out.on, SigF.out.off, Spo0A.out.on, Spo0E.out.on, SigH.out.on, Hpr.out.off, SinR.out.off, SinI.out.off

AbrB is an important "marker" for the regulation of many genes, which is understandable from the Boolean rules with hindsight. By a PubMed query, a direct confirmation was found for the downregulation of SigF together with upregulation of AbrB [C. A. Tomas et al. 2003].

Querying the Stem Base

"For example, we know that in the absence of nutritional stress, sporulation should never be initiated [De 04]. We can use model checking to show this holds in our model by proving that no reachable state exists with SigF = 1 starting from any initial state in which Signal = 0, SigF = 0 and Spo0AP = 0." [Ste07, 341]

This is equivalent to the implication following from the stem base:

$$\text{Signal.in.off, SigF.in.off, Spo0AP.in.off} \rightarrow \text{SigF.out.off}$$

A Query to the PROLOG Knowledge Base

```
:- table off/1.        on(signal.in)  :- on(signal.out).
:- table on/1.         on(signal.out) :- on(signal.in).
off(sigF.in).          on(sigA.in)    :- on(sigA.out).
off(spo0AP.in).        on(sigA.out)   :- on(sigA.in).
on(sigF.out).          off(abrB.out)  :- on(sigF.out).
                       off(kinA.out)  :- on(sigF.out).
   off(spo0AP.out) :- off(spo0A.in), on(sigF.out).
   on(signal)      :- off(sigH.in), on(sigF.out).
```

SigF.in.off Spo0AP.in.off SigF.out.on
→ Signal.in.on Signal.out.on SigA.in.on SigA.out.on Spo0AP.out.off Spo0A.out.off AbrB.out.off KinA.out.off Hpr.out.off

The initial presence of the signal and the ubiquitous transcription factor SigA only are necessary for the initiation of sporulation.

Outline

Towards Realistic Computability

- Attribute exploration is an exponential algorithm.
- Querying the stem base is possible in linear time.
- Background knowledge: state $\rightarrow$ transition $\rightarrow$ transitive context.
- Automatic decision of implications by model checking or automatic text mining.
- Structural and functional analysis of Boolean networks [S. Klamt 2006].
- Conclude dynamical properties of Boolean networks by regarding them as polynomial dynamical systems over finite fields and by exploiting theoretical work in the context of cellular automata [Laubenbacher 2005].

Further Research

- Comparison to observed time series: current work related to the construction and deconstruction of the extracellular matrix in the case of human rheumatoid arthritis. (See also Wollbold 2007.)
- Validation of the simulated transitions: thresholds of support and confidence for the observed transitions.
- Expert based exploration.
- Split attribute exploration into partial problems - decomposition theory of concept lattices.

Literature

Hidde De Jong et al.
Qualitative simulation of the initiation of sporulation in *Bacillus subtilis*.
Bulletin of Mathematical Biology, 66:261–299, 2004.

Bernhard Ganter and Rudolf Wille.
Formal Concept Analysis - Mathematical Foundations.
Springer, Heidelberg, 1999.

L. Jason Steggles et al.
Qualitatively modelling and analysing genetic regulatory networks: a Petri net approach.
Bioinformatics, 23(3):336–343, 2007.