Three Applications of Formal Concept Analysis in Biochemistry and Systems Biology

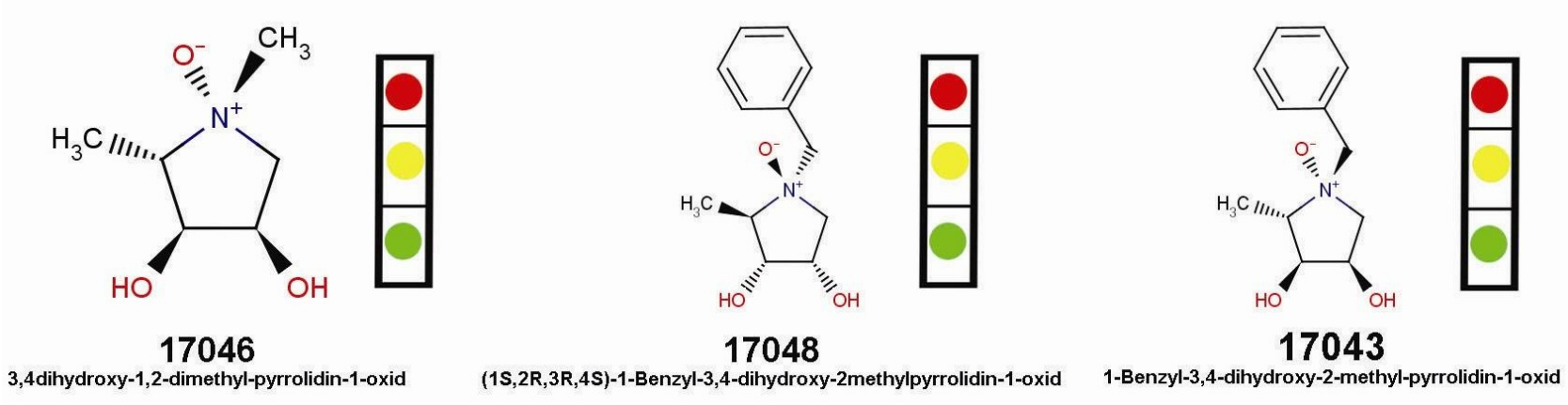
Johannes Wollbold
Hans-Knöll-Institute
Molecular and Applied Microbiology / Systems Biology
Beutenbergstrasse 11a, D-07745 Jena
johannes.wollbold@hki-jena.de

24th May 2007

Outline

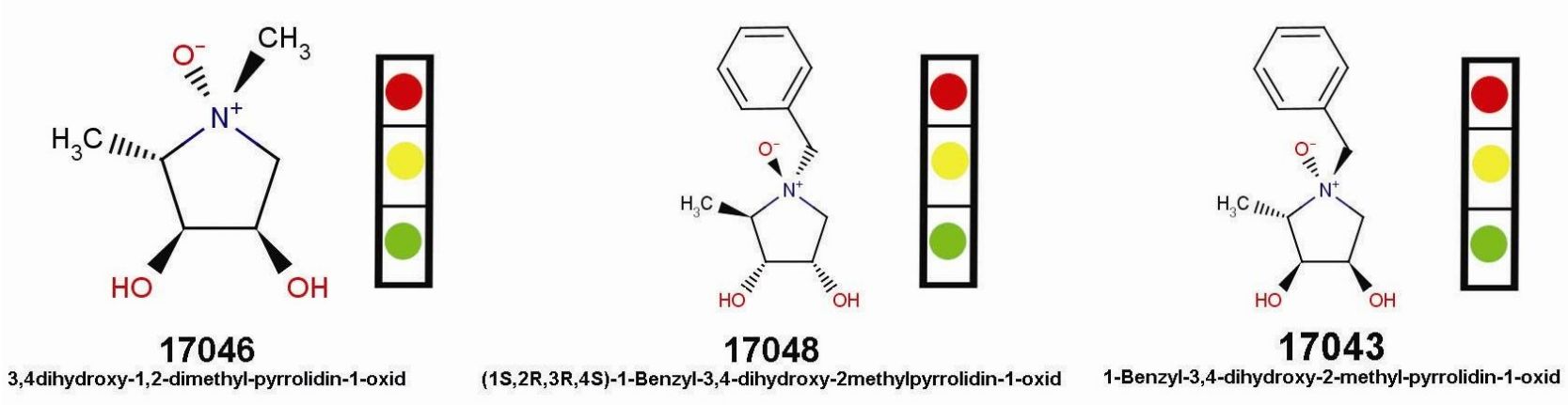
Example: HKI Natural Products Pool

- Collection of about 9.000 natural products, derivatives and synthetic analogues.
- Analysis of a screening for new drugs against tuberculosis: Substances with specific reaction against the M. tuberculosis complexes pknA:atpD ("hit1") or sigB:Rv2050 ("hit3")
- Ranking within classes according to significantly frequent combinations.
- The aim now: Order all substances by their fragments - may indicate structural similarity.



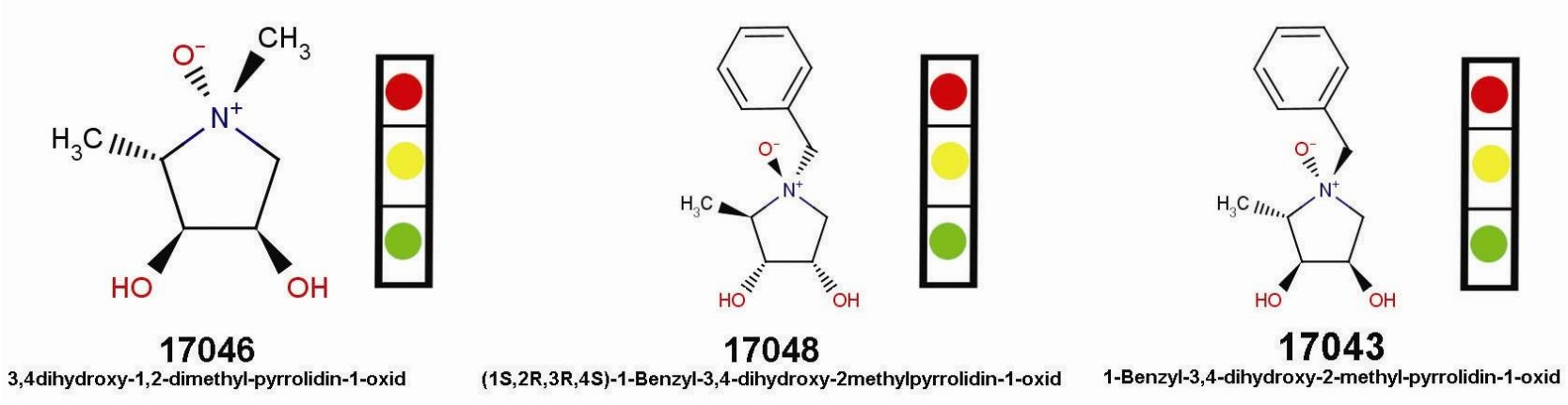
Example: HKI Natural Products Pool

- Collection of about 9.000 natural products, derivatives and synthetic analogues.
- Analysis of a screening for new drugs against tuberculosis: Substances with specific reaction against the M. tuberculosis complexes pknA:atpD ("hit1") or sigB:Rv2050 ("hit3")
- Ranking within classes according to significantly frequent combinations.
- The aim now: Order all substances by their fragments - may indicate structural similarity.



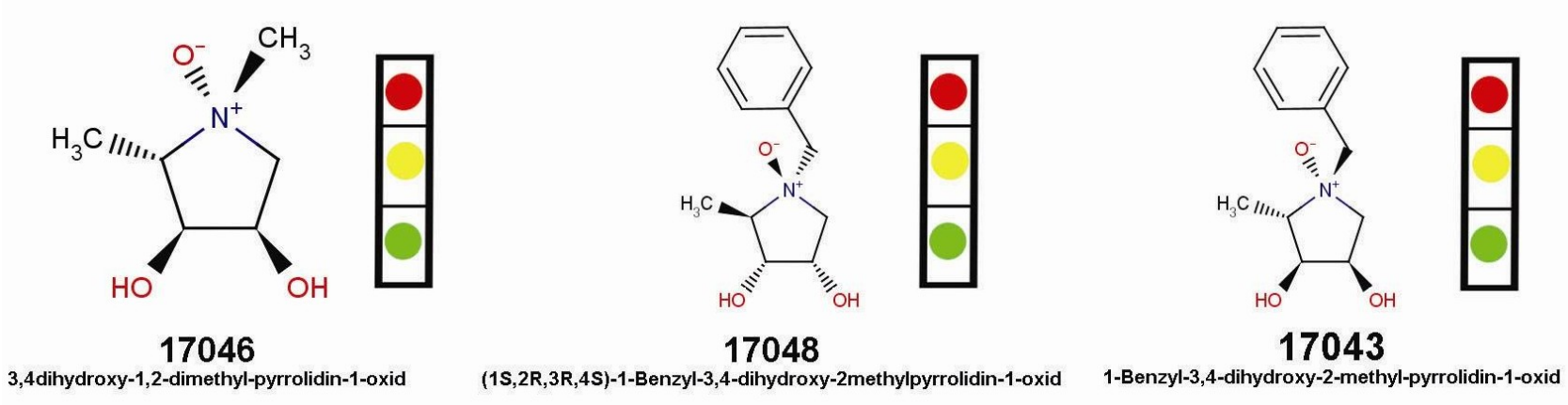
Example: HKI Natural Products Pool

- Collection of about 9.000 natural products, derivatives and synthetic analogues.
- Analysis of a screening for new drugs against tuberculosis: Substances with specific reaction against the M. tuberculosis complexes pknA:atpD ("hit1") or sigB:Rv2050 ("hit3")
- Ranking within classes according to significantly frequent combinations.
- The aim now: Order all substances by their fragments - may indicate structural similarity.



Example: HKI Natural Products Pool

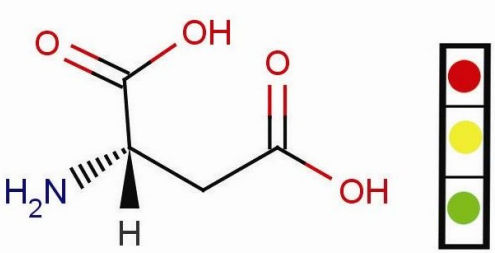
- Collection of about 9.000 natural products, derivatives and synthetic analogues.
- Analysis of a screening for new drugs against tuberculosis: Substances with specific reaction against the M. tuberculosis complexes pknA:atpD ("hit1") or sigB:Rv2050 ("hit3")
- Ranking within classes according to significantly frequent combinations.
- The aim now: Order all substances by their fragments - may indicate structural similarity.

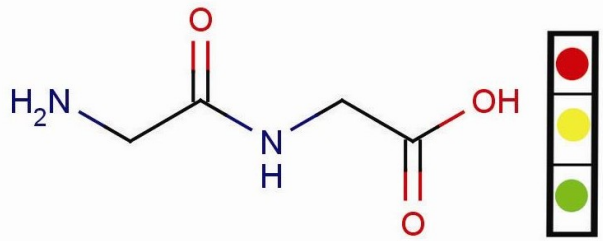


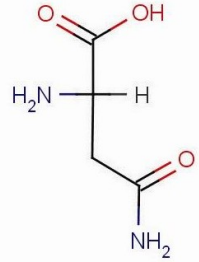
Formal Concepts

fragment	hit1	hit3	2 DOI	3I	5I	8I	9I	6I	7I	4I	6r	7r
substance												
17046	x									x	x	
17043	x									x	x	x
1133		x	x	x	x	x	x					
2892		x	x	x	x	x	x	x				
1055		x	x	x	x	x	x	x	x			

2 fragments common to substances 2892 and 1055:
2DOI C;-O;=O; 3I O;=C;







1133
L-Asparaginsäure

1055
Glycyl-glycin

2892
D/L-Asparagin
(Racemat)

Formal Concepts

fragment	hit1	hit3	2 DOI	3I	5I	8I	9I	6I	7I	4I	6r	7r
substance												
17046	x									x	x	
17043	x									x	x	x
1133		x	x	x	x	x	x					
2892		x	x	x	x	x	x	x				
1055		x	x	x	x	x	x	x	x			

2 fragments common to substances 2892 and 1055:
2DOI C;-O;=O; 3I O;=C;

All fragments common to substances 2892 and 1055: hit3, 2DOI, 3I 0I, 5I, 8I, 9I, 6I

$\Rightarrow$ Formal concept
($\{$2892, 1055$\}$, $\{$hit3, 2DOI, 3I, 5I, 8I, 9I, 6I$\}$)

Formal Concepts

fragment	hit1	hit3	2 DOI	3I	5I	8I	9I	6I	7I	4I	6r	7r
substance												
17046	x									x	x	
17043	x									x	x	x
1133		x	x	x	x	x	x					
2892		x	x	x	x	x	x	x				
1055		x	x	x	x	x	x	x	x			

2 fragments common to substances 2892 and 1055: 2DOI, 3I

Formal Concepts

fragment / substance	hit1	hit3	2 DO I	3I	5I	8I	9I	6I	7I	4I	6r	7r
17046	x									x	x	
17043	x									x	x	x
1133		x	x	x	x	x	x					
2892		x	x	x	x	x	x	x				
1055		x	x	x	x	x	x	x	x			

2 fragments common to substances 2892 and 1055: 2DOI, 3I

All substances with attributes 2DOI and 3I: 1133, 2892 and 1055

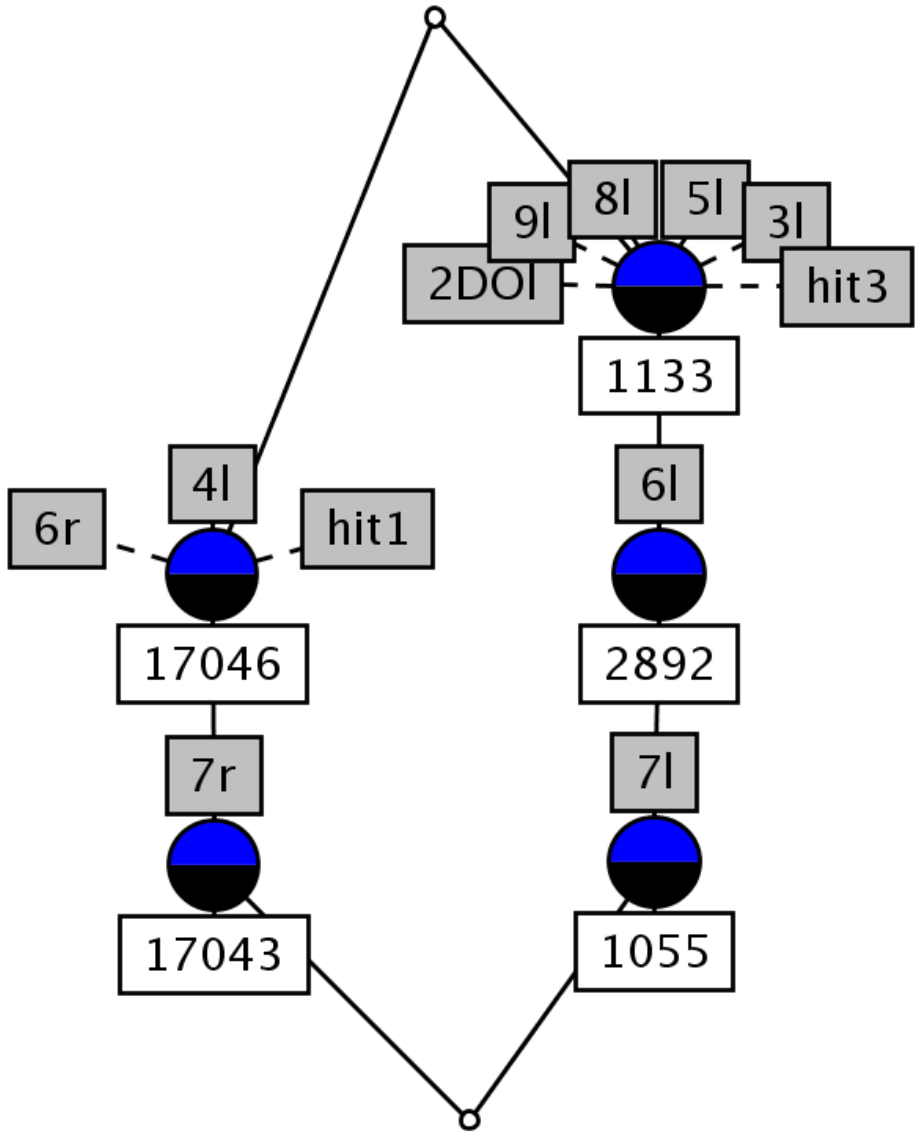
Formal Concepts

fragment	hit1	hit3	2DOI	3l	5l	8l	9l	6l	7l	4l	6r	7r
substance												
17046	x									x	x	
17043	x									x	x	x
1133		x	x	x	x	x	x					
2892		x	x	x	x	x	x	x				
1055		x	x	x	x	x	x	x	x			

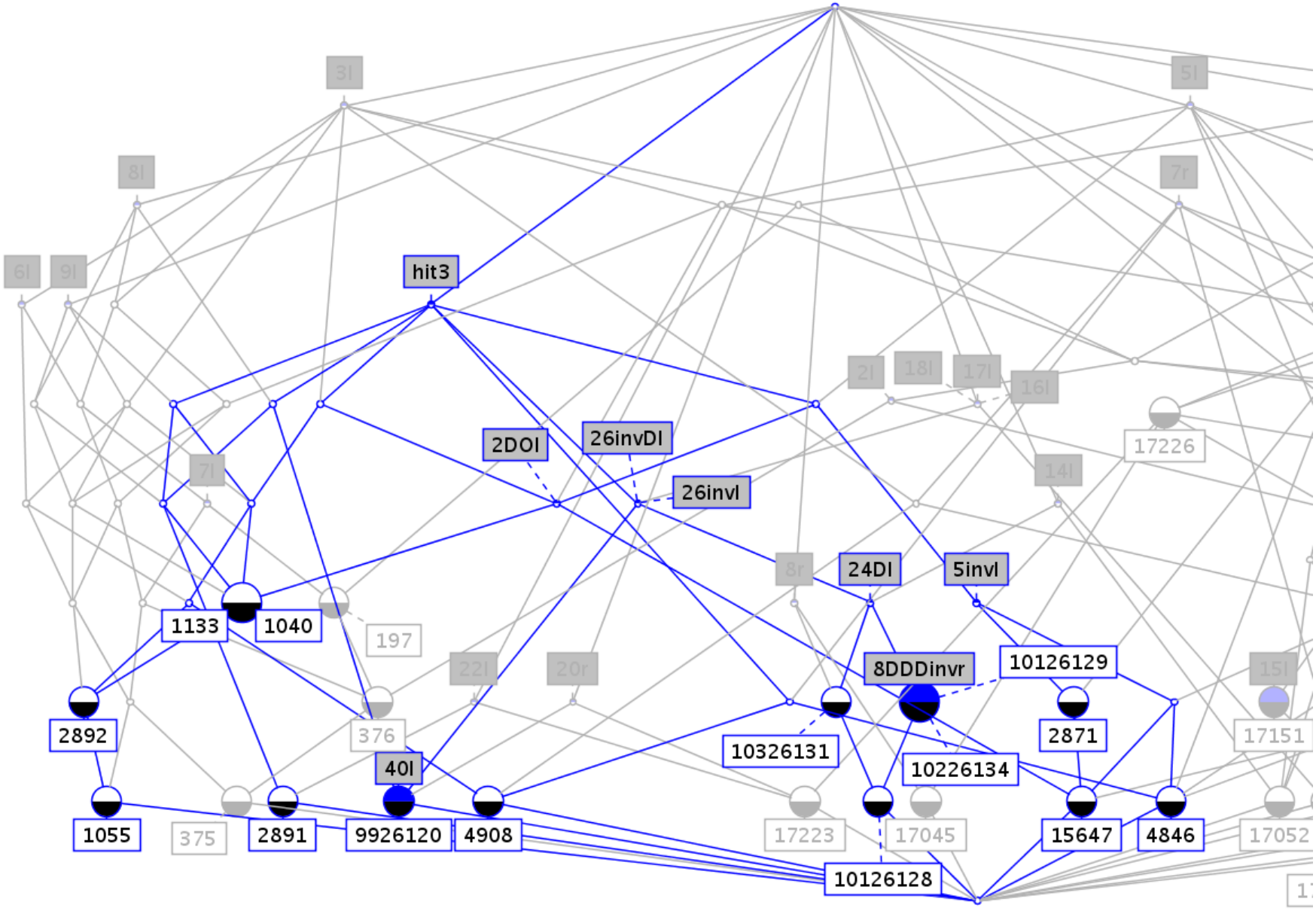
2 fragments common to substances 2892 and 1055: 2DOI, 3l
All substances with attributes 2DOI and 3l: 1133, 2892 and 1055
All attributes common to substances 1133, 2892 and 1055: hit3, 2DOI, 3l, 5l, 8l, 9l

$\Rightarrow$ Formal concept
($\{$1133, 2892, 1055$\}$, $\{$hit3, 2DOI, 3l, 5l, 8l, 9l$\}$)

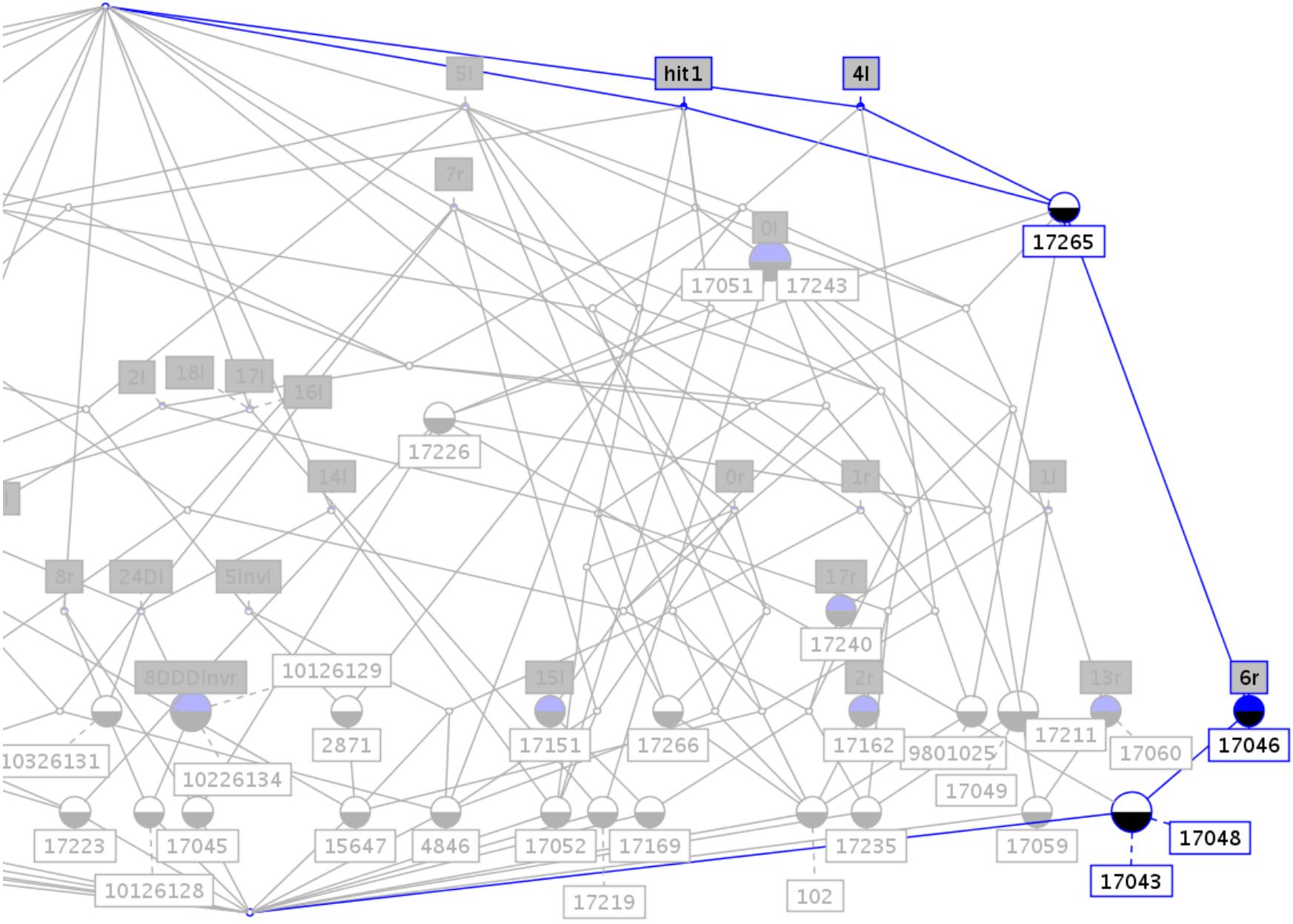
The Conceptual Hierarchy



The Class of Substances Specific Against sigB:Rv2050

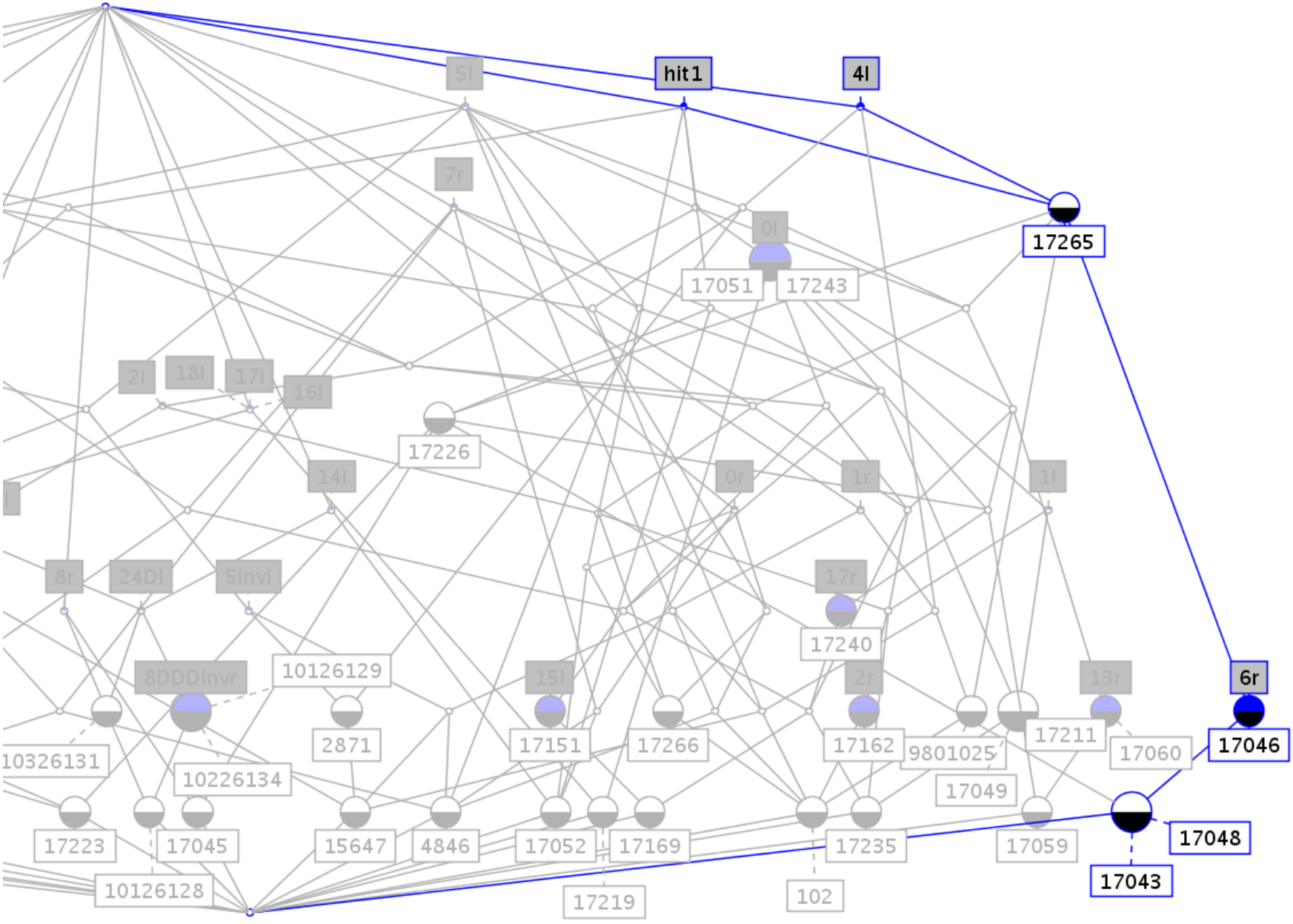


A Fragment Specific Against pknA:atpD?



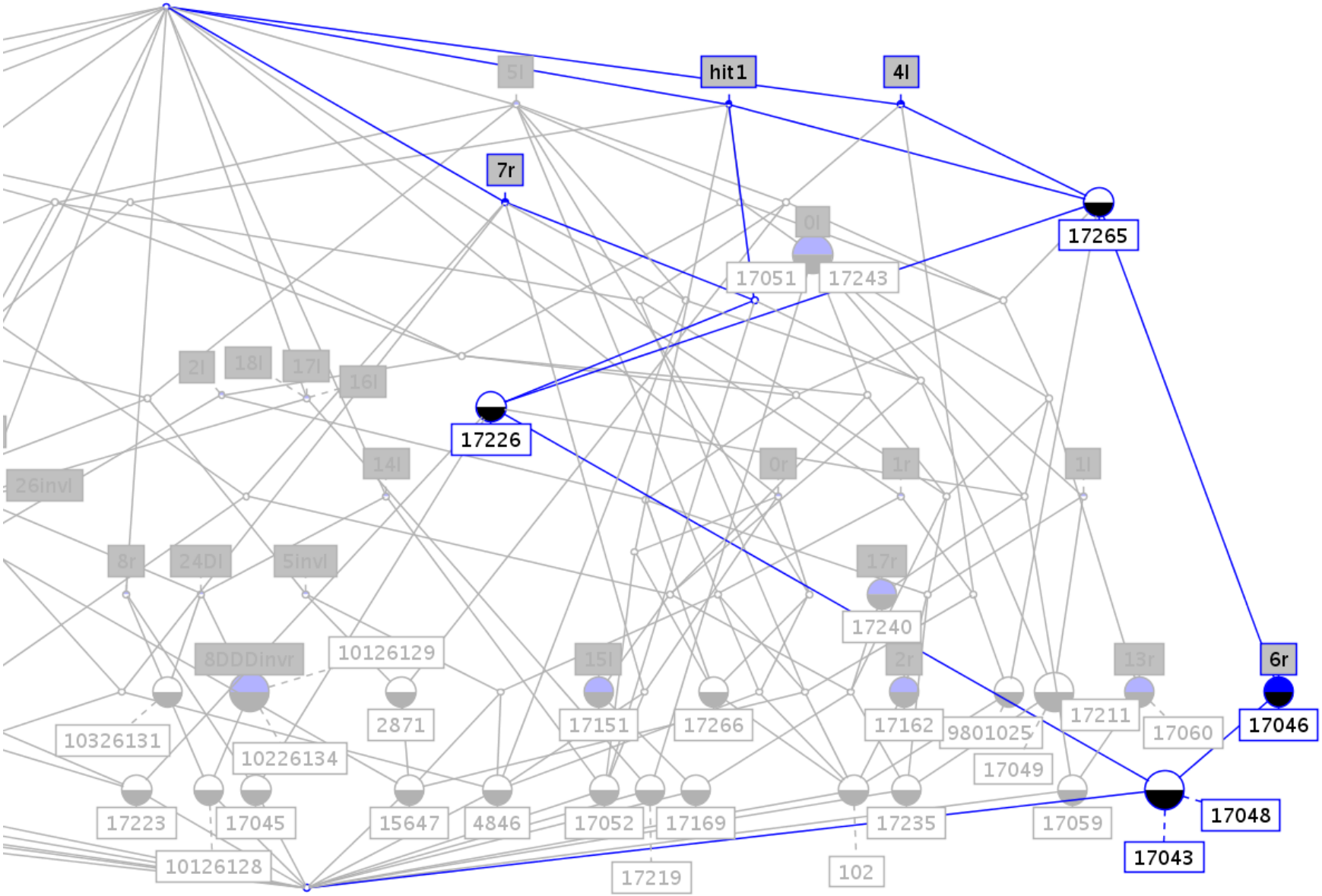
6r $\longrightarrow$ 4l, hit1.

A Fragment Specific Against pknA:atpD?

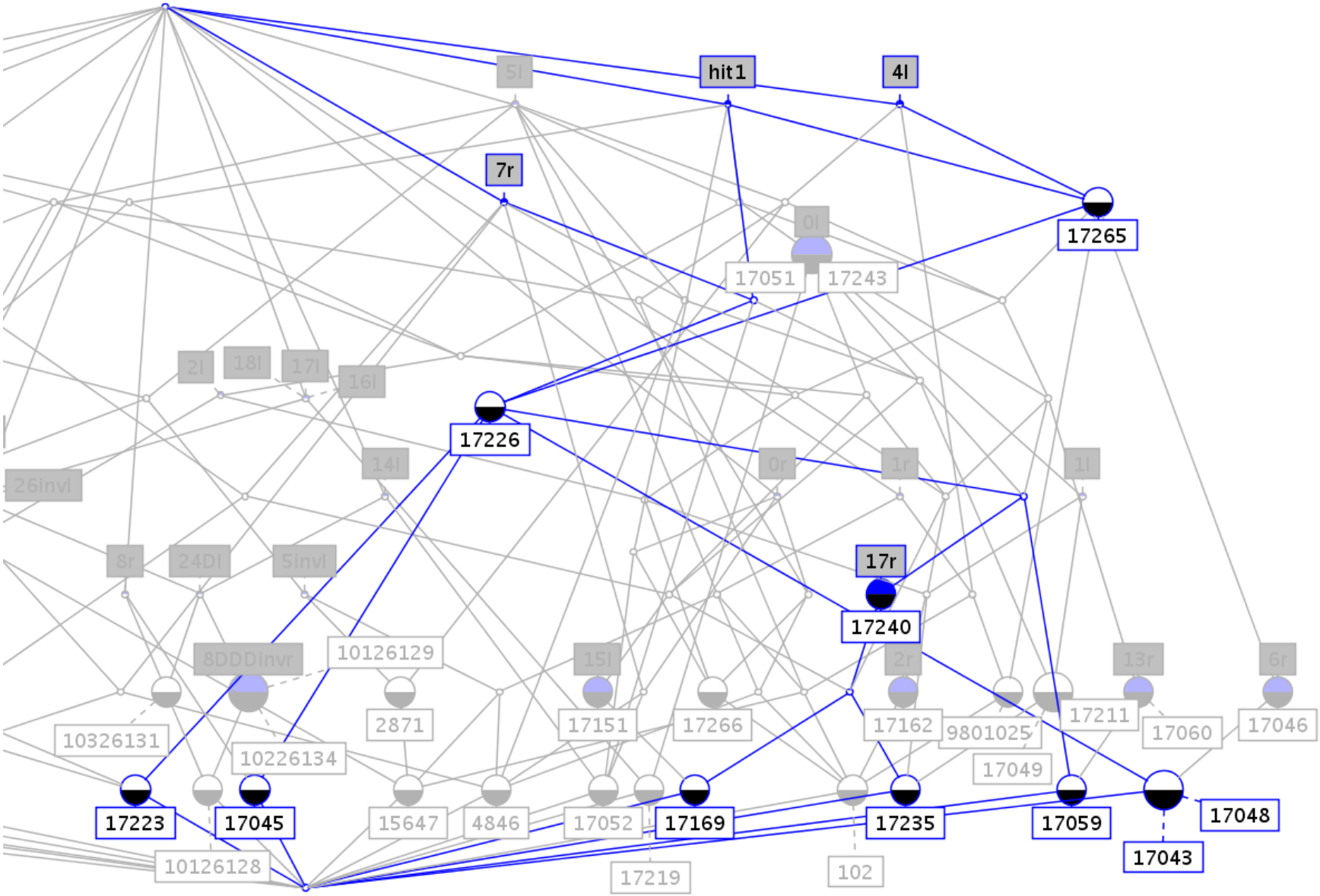


6r $\rightarrow$ 4l, hit1.

A Fragment Combination Specific Against pknA:atpD?



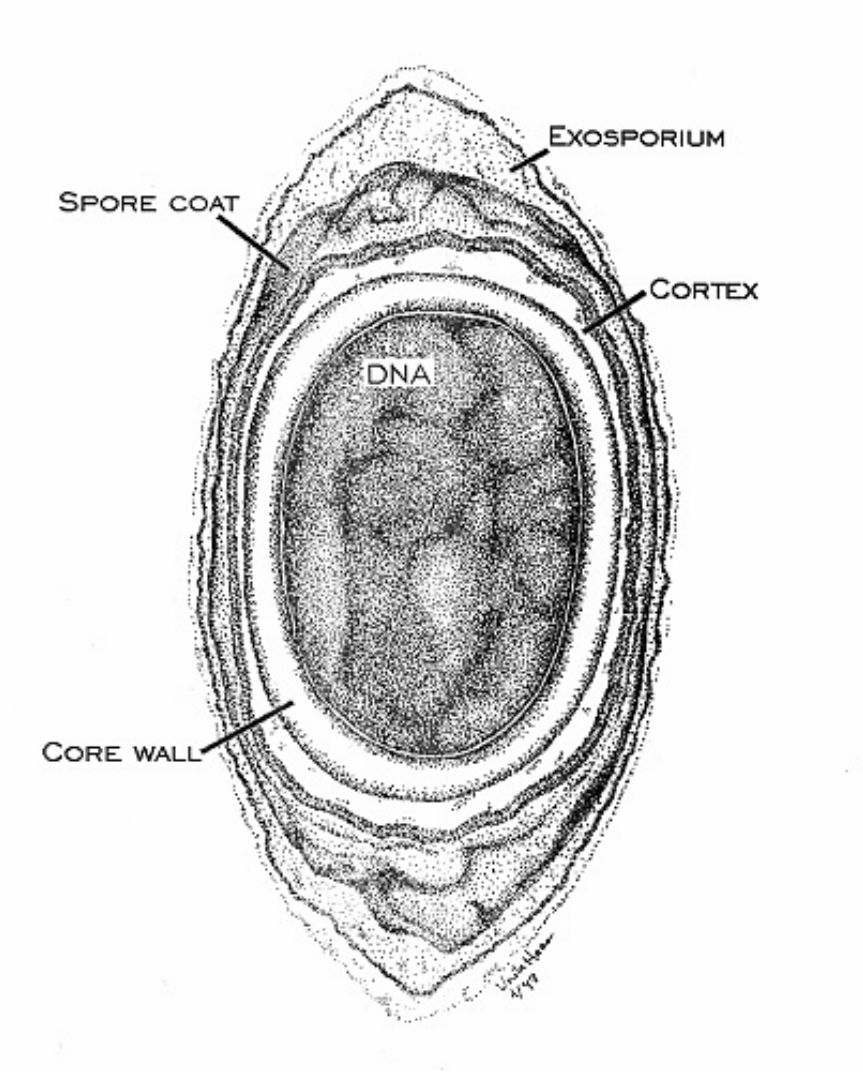
A Fragment Combination Specific Against pknA:atpD



4l, 7r $\longrightarrow$ hit1

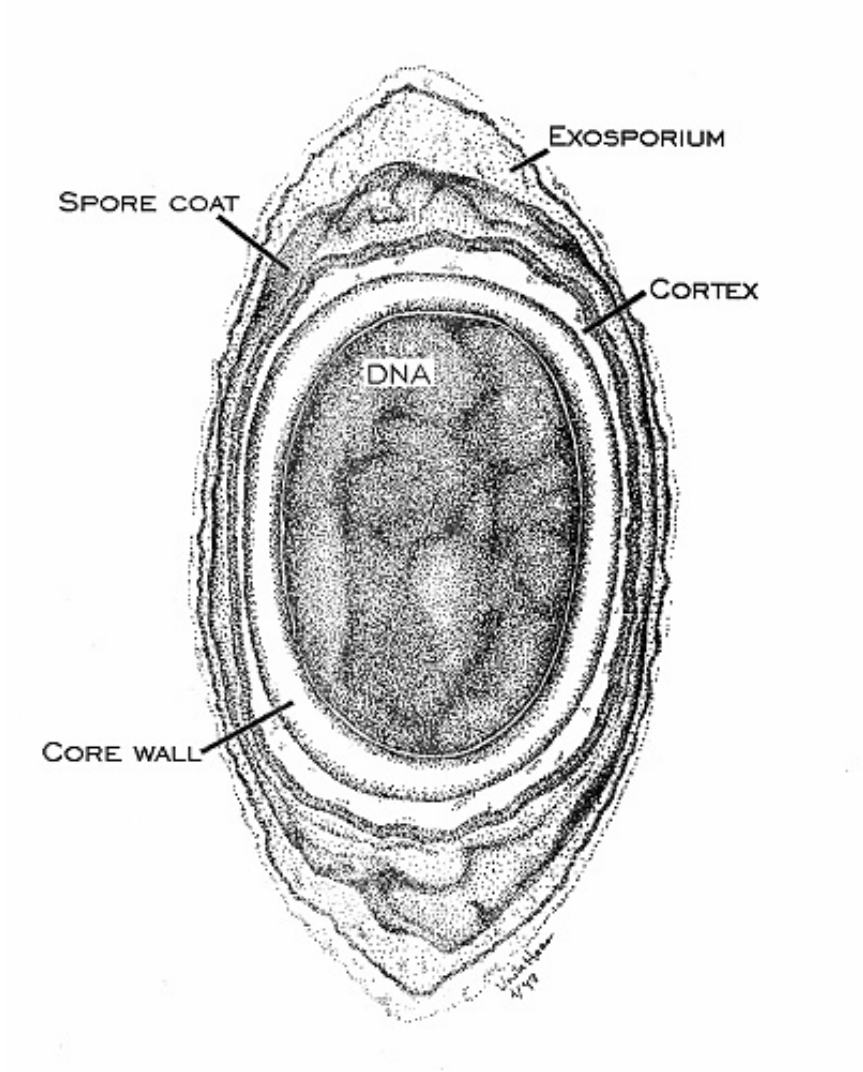
Outline

Bacillus subtilis



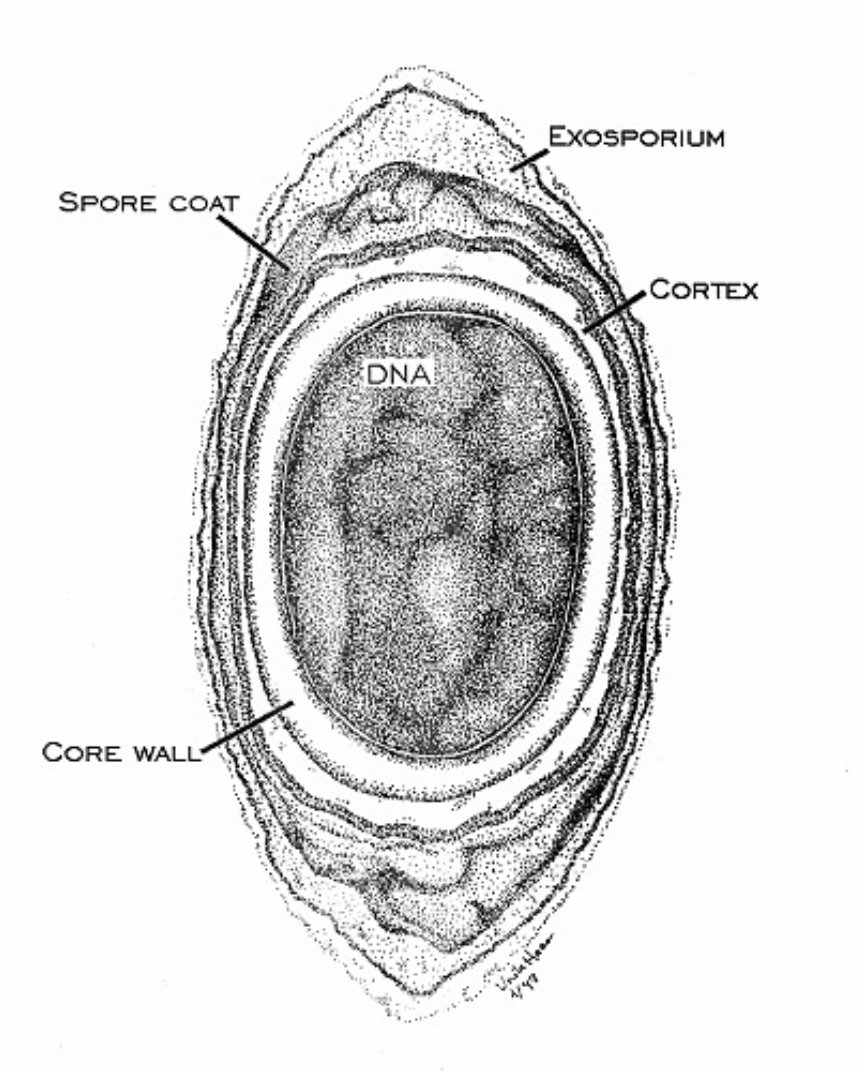
- Gram positive soil bacterium.
- Produces single endospores under environmental stress, which can survive ultraviolet or gamma radiation, acid, hours of boiling, or starvation.
- Switch between two completely different genetic programs.
- Model organism.

Bacillus subtilis



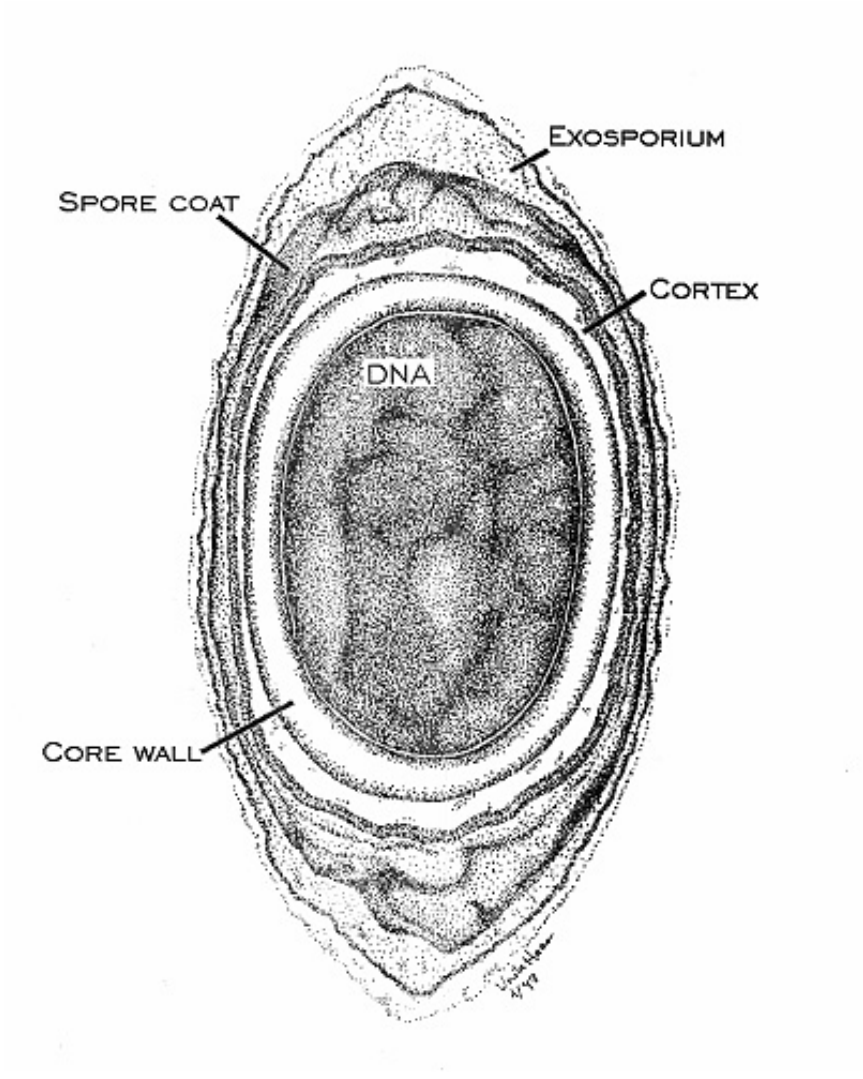
- Gram positive soil bacterium.
- Produces single endospores under environmental stress, which can survive ultraviolet or gamma radiation, acid, hours of boiling, or starvation.
- Switch between two completely different genetic programs.
- Model organism.

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.

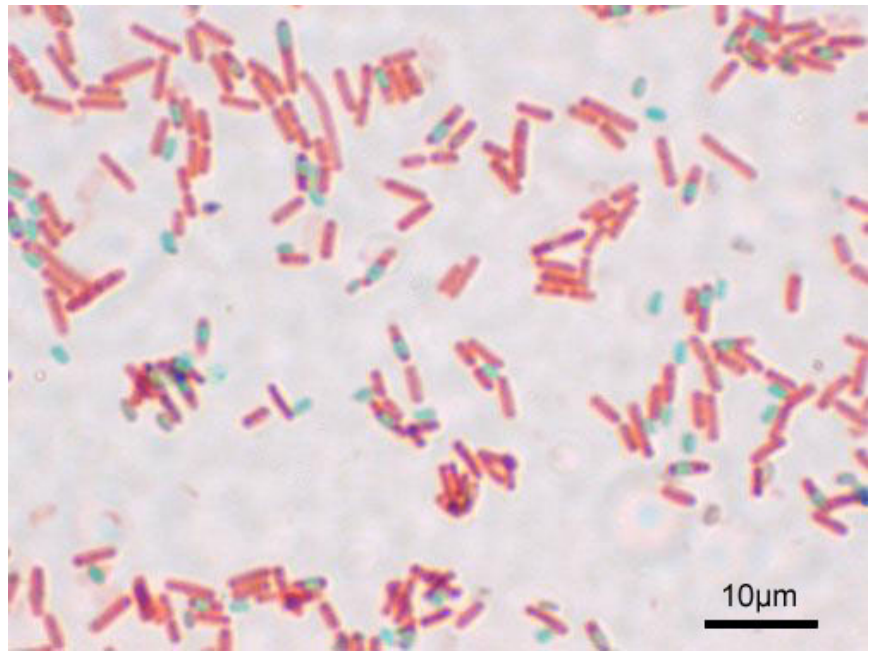
Bacillus subtilis



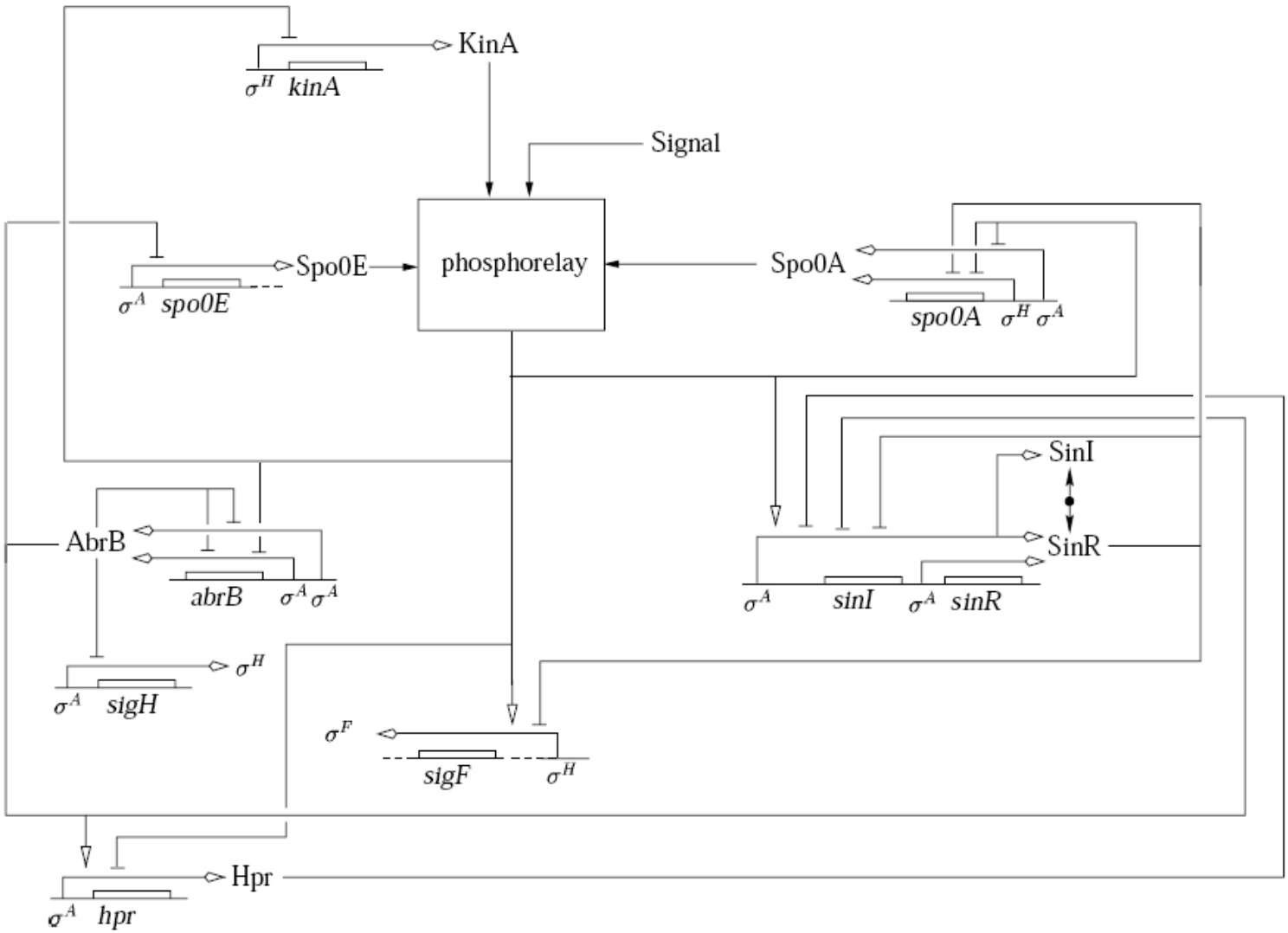
- Gram positive soil bacterium.
- Produces single endospores under environmental stress, which can survive ultraviolet or gamma radiation, acid, hours of boiling, or starvation.
- Switch between two completely different genetic programs.
- Model organism.

Some Main Regulators

- Signal: nutritional stress.
- Phosphorylation of the transcriptional regulator Spo0A by KinA, Spo0E (phosphatase) and Signal.
- Sigma transcription factors σ^A, σ^F and σ^H.
- Transcription regulators AbrB and Hpr.



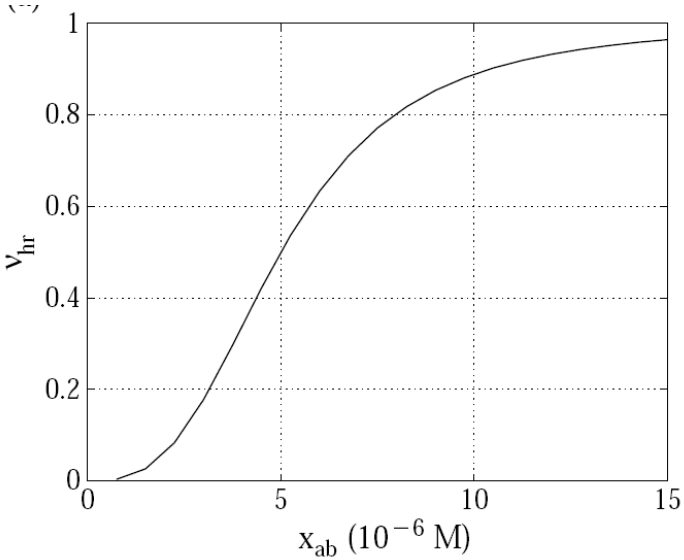
A Gene Regulatory Network

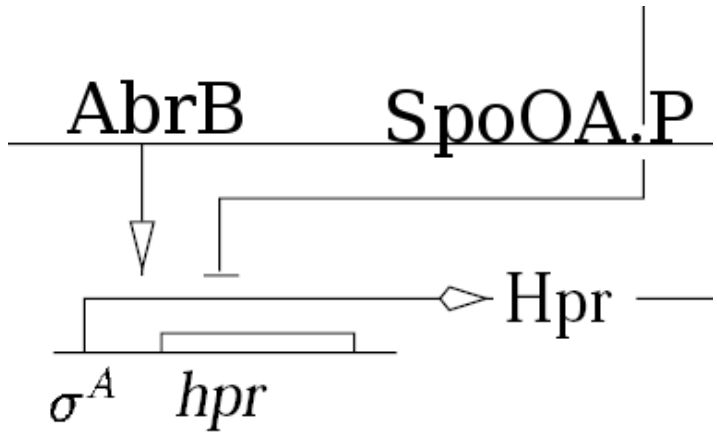


[Hidde de Jong. Qualitative Simulation of the Initiation of Sporulation in Bacillus subtilis. In: *Bulletin of Mathematical Biology* (2004) **66**, 261-299.]

Modeling the Dynamics as a Boolean Network

Thresholds for gene expression $\Rightarrow$ 0/1 levels capture essential properties of the network. Simultaneous update at discrete time points according to Boolean functions [Kau95]:





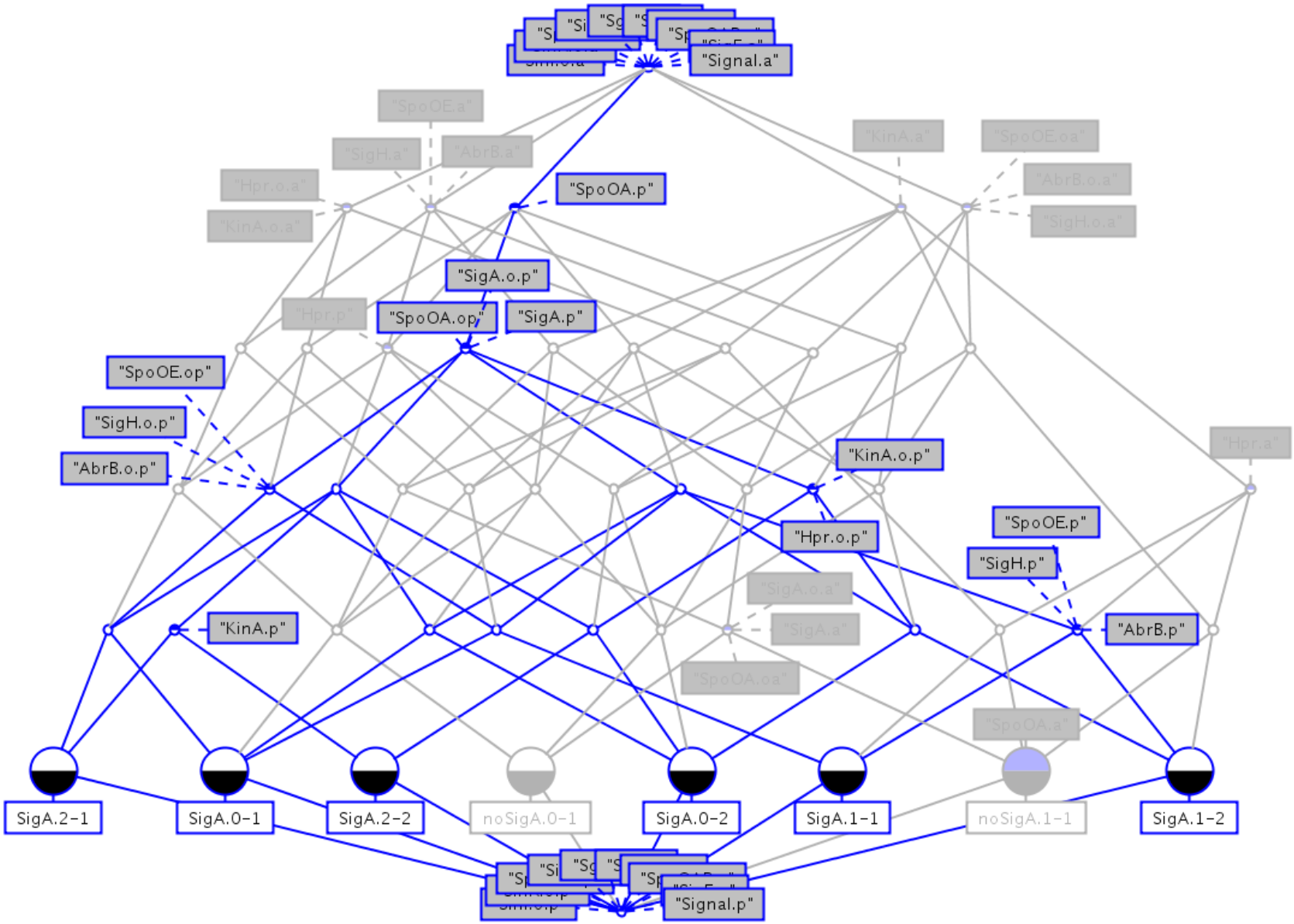
$$\text{Hpr.out} = \text{SigA.in} \wedge \text{AbrB.in} \wedge \overline{\text{SpoOA.P.in}}$$

[L. J. Steggles et al. Qualitatively modelling and analysing genetic regulatory networks: a Petri net approach. In: *Bioinformatics* 23 (3), 2007, 336-343.]

Simulations Starting from an Initial State without Stress

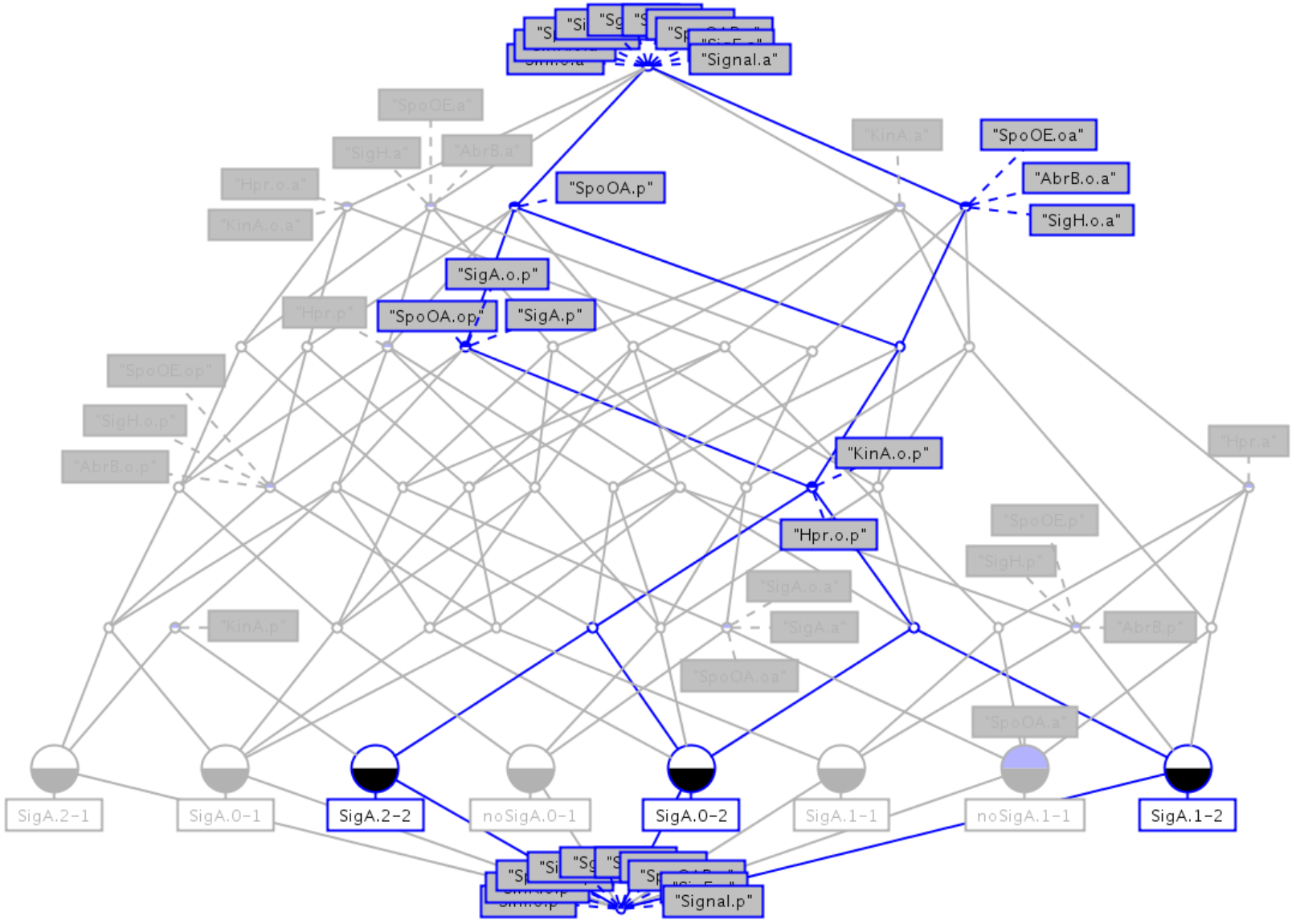
Entity / State	SigA.0	SigA.1	SigA.2	SigA.1	noSigA.0	noSigA.1	noSigA.1
Signal	0	0	0	0	0	0	0
SigA	1	1	1	1	0	0	0
SigF	0	0	0	0	0	0	0
KinA	0	0	1	0	0	0	0
SpoOA	1	1	1	1	1	0	0
SpoOA.P	0	0	0	0	0	0	0
AbrB	0	1	0	1	0	0	0
SpoOE	0	1	0	1	0	0	0
SigH	0	1	0	1	0	0	0
Hpr	1	0	1	0	1	0	0
SinR	0	0	0	0	0	0	0
SinI	0	0	0	0	0	0	0

Transitions with SpoOA.out.pres



$< 6 >$ SpoOA.out.pres $\rightarrow$ SigA.in.pres SpoOA.in.pres SigA.out.pres

Transitions with Hpr.out.pres



$< 3 >$ Hpr.out.pres $\rightarrow$ KinA.out.pres, SpoOA.out.pres, AbrB.out.abs, SpoOE.out.abs, SigH.out.abs

Outline

Attribute Exploration

- Interactive algorithm.
- Generates rules $A \longrightarrow B$ between attribute sets of a given data table.
- An expert or a computer program decides about the general validity of the rule.
- No $\curvearrowright$ counterexample.
- Yes $\curvearrowright$ add $A \longrightarrow B$ to the stem base of the context (sound, complete, non redundant).
- All rules valid in the explored domain are derivable from this knowledge base, e.g. by PROLOG.

Attribute Exploration

- Interactive algorithm.
- Generates rules $A \longrightarrow B$ between attribute sets of a given data table.
- An expert or a computer program decides about the general validity of the rule.
- No $\curvearrowright$ counterexample.
- Yes $\curvearrowright$ add $A \longrightarrow B$ to the stem base of the context (sound, complete, non redundant).
- All rules valid in the explored domain are derivable from this knowledge base, e.g. by PROLOG.

Attribute Exploration

- Interactive algorithm.
- Generates rules $A \longrightarrow B$ between attribute sets of a given data table.
- An expert or a computer program decides about the general validity of the rule.
- No $\curvearrowright$ counterexample.
- Yes $\curvearrowright$ add $A \longrightarrow B$ to the stem base of the context (sound, complete, non redundant).
- All rules valid in the explored domain are derivable from this knowledge base, e.g. by PROLOG.

Attribute Exploration

- Interactive algorithm.
- Generates rules $A \longrightarrow B$ between attribute sets of a given data table.
- An expert or a computer program decides about the general validity of the rule.
- No $\curvearrowright$ counterexample.
- Yes $\curvearrowright$ add $A \longrightarrow B$ to the stem base of the context (sound, complete, non redundant).
- All rules valid in the explored domain are derivable from this knowledge base, e.g. by PROLOG.

Attribute Exploration

- Interactive algorithm.
- Generates rules $A \longrightarrow B$ between attribute sets of a given data table.
- An expert or a computer program decides about the general validity of the rule.
- No $\curvearrowright$ counterexample.
- Yes $\curvearrowright$ add $A \longrightarrow B$ to the stem base of the context (sound, complete, non redundant).
- All rules valid in the explored domain are derivable from this knowledge base, e.g. by PROLOG.

Attribute Exploration

- Interactive algorithm.
- Generates rules $A \longrightarrow B$ between attribute sets of a given data table.
- An expert or a computer program decides about the general validity of the rule.
- No $\curvearrowright$ counterexample.
- Yes $\curvearrowright$ add $A \longrightarrow B$ to the stem base of the context (sound, complete, non redundant).
- All rules valid in the explored domain are derivable from this knowledge base, e.g. by PROLOG.

Example: Inspection of all Possible Transitions in the B. subtilis Network

- From all possible $2^{12} = 4096$ initial states, 4224 transitions[1] were generated (quite deterministic rules).
- 524 implications (with support > 0) in the stem base, but $11.023.494 \approx 2^{24}$ concepts.
- Model validation according to [Sea07, 341f.], during attribute exploration or by PROLOG queries to the stem base.

[1] and 11.700 transitive transitions

Verification of Some Implications in the Experimental Literature.

Signal.abs, SigF.abs, SpoOA.P.abs $\longrightarrow$ SigF.out.abs

In the absence of nutritional stress sporulation should never be initiated [de Jong et al., 2004].

SigF.out.pres $\longrightarrow$ KinA.out.abs, SpoOA.out.abs, Hpr.out.abs, AbrB.out.abs

Mutual exclusion: SpoOA.P is reported to activate the production of SigF but also repress its own production. [de Jong et al., 2004]

SigH.out.abs $\longrightarrow$ AbrB.out.abs, SpoOE.out.abs, SinR.out.abs, SinI.out.abs

All these genes are regulated $\overline{gene.out} = \overline{\text{SigA.in}} \wedge \text{AbrB.in} \ (\wedge \ ...)$. Such dependencies can be checked systematically.

Verification of Some Implications in the Experimental Literature.

Signal.abs, SigF.abs, SpoOA.P.abs $\longrightarrow$ SigF.out.abs

In the absence of nutritional stress sporulation should never be initiated [de Jong et al., 2004].

SigF.out.pres $\longrightarrow$ KinA.out.abs, SpoOA.out.abs, Hpr.out.abs, AbrB.out.abs

Mutual exclusion: SpoOA.P is reported to activate the production of SigF but also repress its own production. [de Jong et al., 2004]

SigH.out.abs $\longrightarrow$ AbrB.out.abs, SpoOE.out.abs, SinR.out.abs, SinI.out.abs

All these genes are regulated $\overline{gene.out} = \overline{\text{SigA.in}} \wedge \text{AbrB.in} \ (\wedge \ ...)$. Such dependencies can be checked systematically.

Verification of Some Implications in the Experimental Literature.

Signal.abs, SigF.abs, SpoOA.P.abs $\longrightarrow$ SigF.out.abs

In the absence of nutritional stress sporulation should never be initiated [de Jong et al., 2004].

SigF.out.pres $\longrightarrow$ KinA.out.abs, SpoOA.out.abs, Hpr.out.abs, AbrB.out.abs

Mutual exclusion: SpoOA.P is reported to activate the production of SigF but also repress its own production. [de Jong et al., 2004]

SigH.out.abs $\longrightarrow$ AbrB.out.abs, SpoOE.out.abs, SinR.out.abs, SinI.out.abs

All these genes are regulated $\overline{gene.out} = \overline{\text{SigA.in}} \wedge \text{AbrB.in}\ (\wedge\ \ldots)$. Such dependencies can be checked systematically.

Supplementary Results

The stem base contains the complete implicational logic of the explored temporal transitions. It may be queried for interesting and simple rules:

$< 4500 >$ SpoOA.P.pres, KinA.out.abs $\longrightarrow$ Hpr.out.abs;
$< 4212 >$ SigH.pres. KinA.out.abs $\longrightarrow$ Hpr.out.abs;
$< 3972 >$ AbrB.abs, KinA.out.abs $\longrightarrow$ Hpr.out.abs:

$\overline{\text{Hpr}}$ and $\overline{\text{KinA}}$ are determined by different Boolean functions, but they are coregulated in many cases.

$< 3904 >$ AbrB.out.pres $\longrightarrow$ SigA.pres, SigA.out.pres, SigF.out.abs, SpoOA.out.pres, SpoOE.out.pres, SigH.out.pres, Hpr.out.abs, SinR.out.abs, SinI.out.abs

AbrB is an important "marker" for the regulation of many genes, since its expression is determined by the simultaneous presence / absence of the 3 main genes SigA, $\overline{\text{AbrB}}$ and $\overline{\text{SpoOA.P}}$, which often have an exclusive (OR-) influence.

Supplementary Results

The stem base contains the complete implicational logic of the explored temporal transitions. It may be queried for interesting and simple rules:

$< 4500 >$ SpoOA.P.pres, KinA.out.abs $\longrightarrow$ Hpr.out.abs;
$< 4212 >$ SigH.pres. KinA.out.abs $\longrightarrow$ Hpr.out.abs;
$< 3972 >$ AbrB.abs, KinA.out.abs $\longrightarrow$ Hpr.out.abs:

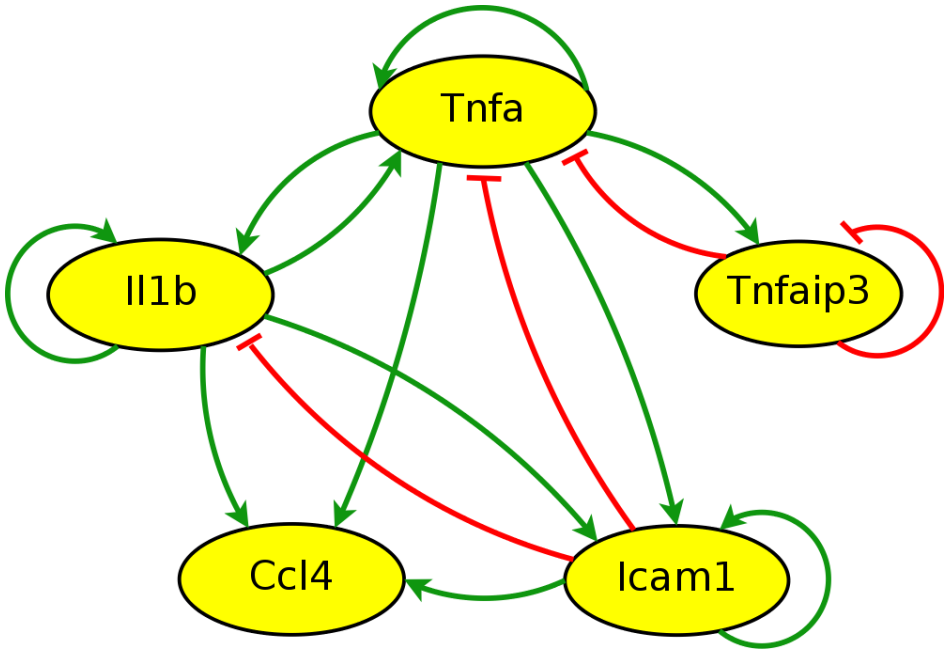
$\overline{\text{Hpr}}$ and $\overline{\text{KinA}}$ are determined by different Boolean functions, but they are coregulated in many cases.

$< 3904 >$ AbrB.out.pres $\longrightarrow$ SigA.pres, SigA.out.pres, SigF.out.abs, SpoOA.out.pres, SpoOE.out.pres, SigH.out.pres, Hpr.out.abs, SinR.out.abs, SinI.out.abs

AbrB is an important "marker" for the regulation of many genes, since its expression is determined by the simultaneous presence / absence of the 3 main genes SigA, $\overline{\text{AbrB}}$ and $\overline{\text{SpoOA.P}}$, which often have an exclusive (OR-) influence.

Outline

A Network of 5 Important Genes



- Simulation starting from the initial state of an observed time series (only Tnfa protein was present).
- Non deterministic generalization in the case of inhibition: If Tnfa.in.pres and Tnfaip3.in.pres, generate 2 output states with Tnfaip3.out.abs or Tnfaip3.out.pres.

Exploration Results for the Simulation, Validated Against the Observation

Il1β.pres^{in} $\longrightarrow$ Icam1.pres^{in} Ccl4.pres^{in}
Ccl4.abs^{in} $\longrightarrow$ ... Icam1.abs^{in} Il1β.abs^{in}

Similar regulation of Il1β, Icam1 and Ccl4.

Il1β.abs^{in} Tnfaip3.pres^{out} $\longrightarrow$ Tnfα.pres^{in}
Tnfα.abs^{in} Tnfaip3.pres^{out} $\longrightarrow$ Il1β.pres^{in}

Mirroring the important role of the upregulating genes Il1β and Tnfα.

Tnfα.abs^{in} Il1β.abs^{in} $\longrightarrow$ Tnfα.abs^{out} Tnfaip3.abs^{out} Il1β.abs^{out}

Structural deadlock in Petri net theory.

$\top \longrightarrow$ Icam1.pres^{out}

$\Leftrightarrow \top \longrightarrow$ always(Icam1.pres^{out}) in $\mathbb{K}_s$.

Exploration Results for the Simulation, Validated Against the Observation

$\text{Il1}\beta.\text{pres}^{in} \longrightarrow \text{Icam1.pres}^{in}\ \text{Ccl4.pres}^{in}$
$\text{Ccl4.abs}^{in} \longrightarrow \ldots\ \text{Icam1.abs}^{in}\ \text{Il1}\beta.\text{abs}^{in}$

Similar regulation of Il1β, Icam1 and Ccl4.

$\text{Il1}\beta.\text{abs}^{in}\ \text{Tnfaip3.pres}^{out} \longrightarrow \text{Tnf}\alpha.\text{pres}^{in}$
$\text{Tnf}\alpha.\text{abs}^{in}\ \text{Tnfaip3.pres}^{out} \longrightarrow \text{Il1}\beta.\text{pres}^{in}$

Mirroring the important role of the upregulating genes Il1β and Tnfα.

$\text{Tnf}\alpha.\text{abs}^{in}\ \text{Il1}\beta.\text{abs}^{in} \longrightarrow \text{Tnf}\alpha.\text{abs}^{out}\ \text{Tnfaip3.abs}^{out}\ \text{Il1}\beta.\text{abs}^{out}$

Structural deadlock in Petri net theory.

$\top \longrightarrow \text{Icam1.pres}^{out}$
$\Leftrightarrow \top \longrightarrow \text{always}(\text{Icam1.pres}^{out})\ \text{in}\ \mathbb{K}_s.$

Exploration Results for the Simulation, Validated Against the Observation

$\text{Il1}\beta.\text{pres}^{in} \rightarrow \text{Icam1.pres}^{in}\ \text{Ccl4.pres}^{in}$
$\text{Ccl4.abs}^{in} \rightarrow \ldots\ \text{Icam1.abs}^{in}\ \text{Il1}\beta.\text{abs}^{in}$

Similar regulation of Il1β, Icam1 and Ccl4.

$\text{Il1}\beta.\text{abs}^{in}\ \text{Tnfaip3.pres}^{out} \rightarrow \text{Tnf}\alpha.\text{pres}^{in}$
$\text{Tnf}\alpha.\text{abs}^{in}\ \text{Tnfaip3.pres}^{out} \rightarrow \text{Il1}\beta.\text{pres}^{in}$

Mirroring the important role of the upregulating genes Il1β and Tnfα.

$\text{Tnf}\alpha.\text{abs}^{in}\ \text{Il1}\beta.\text{abs}^{in} \rightarrow \text{Tnf}\alpha.\text{abs}^{out}\ \text{Tnfaip3.abs}^{out}\ \text{Il1}\beta.\text{abs}^{out}$

Structural deadlock in Petri net theory.

$\top \rightarrow \text{Icam1.pres}^{out}$
$\Leftrightarrow \top \rightarrow \text{always}(\text{Icam1.pres}^{out})$ in $\mathbb{K}_s$.

Exploration Results for the Simulation, Validated Against the Observation

$\text{Il1}\beta.\text{pres}^{in} \rightarrow \text{Icam1.pres}^{in}\ \text{Ccl4.pres}^{in}$
$\text{Ccl4.abs}^{in} \rightarrow \ldots\ \text{Icam1.abs}^{in}\ \text{Il1}\beta.\text{abs}^{in}$

Similar regulation of Il1β, Icam1 and Ccl4.

$\text{Il1}\beta.\text{abs}^{in}\ \text{Tnfaip3.pres}^{out} \rightarrow \text{Tnf}\alpha.\text{pres}^{in}$
$\text{Tnf}\alpha.\text{abs}^{in}\ \text{Tnfaip3.pres}^{out} \rightarrow \text{Il1}\beta.\text{pres}^{in}$

Mirroring the important role of the upregulating genes Il1β and Tnfα.

$\text{Tnf}\alpha.\text{abs}^{in}\ \text{Il1}\beta.\text{abs}^{in} \rightarrow \text{Tnf}\alpha.\text{abs}^{out}\ \text{Tnfaip3.abs}^{out}\ \text{Il1}\beta.\text{abs}^{out}$

Structural deadlock in Petri net theory.

$\top \rightarrow \text{Icam1.pres}^{out}$

$\Leftrightarrow \top \rightarrow \text{always}(\text{Icam1.pres}^{out})$ in $\mathbb{K}_s$.

Outline

Open Problems

- Solve mathematical and logical questions, such that attribute exploration may be applied to larger networks.
- Definition of the expert's role. New transition as counterexample $\longrightarrow$ new transition rule, check affected previous implications?
- Develop a biologically more exact, comprehensive and realistic model (finer steps, cell type...).
- Application to proteins related to the extracellular matrix.

Acknowledgements

Axel Brakhage, HKI Jena
Reinhard Guthke, HKI Jena
Christian Hummert, HKI Jena
Dörte Radke, HKI Jena

Ute Möllmann, HKI Jena
Lennart Heinzerling, University Saarbrücken
Maik Friedel, FLI Jena
Ulrike Gaussmann, FLI Jena
René Huber, FSU Jena / Experimental Rheumatology

Bernhard Ganter, TU Dresden

... and thanks for your attention!

Acknowledgements

Axel Brakhage, HKI Jena
Reinhard Guthke, HKI Jena
Christian Hummert, HKI Jena
Dörte Radke, HKI Jena

Ute Möllmann, HKI Jena
Lennart Heinzerling, University Saarbrücken
Maik Friedel, FLI Jena
Ulrike Gaussmann, FLI Jena
René Huber, FSU Jena / Experimental Rheumatology

Bernhard Ganter, TU Dresden

... and thanks for your attention!

Literature


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


S.A. Kauffman.
At home in the universe: the search for laws of self-organization and complexity.
Oxford University Press, New York, 1995.


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


S. Yevtushenko.
Concept Explorer Java tool.
http://sourceforge.net/projects/conexp, 2003.