

Relevance of Information in Cell Signaling Pathways using Default Logic

Andrei Doncescu¹, Pierre Siegel², Tan. Le¹

¹LAAS-CNRS, University of Toulouse, Toulouse, France,

²Aix Marseille Universite, CNRS, LIF UMR 7279, 13288, Marseille, France

- Cell Signaling Pathway Simulation is a very useful tool in the drug discovery process.
- These simulation programs can be divided into :
 - Dynamic simulation, based on differential equations and could be considered in "real-time",
 - Knowledge-Based Discovery (KDB) Program, the consistency of the model is checked.
- The most efficient KBDP approach is based on first order logic (FOL).

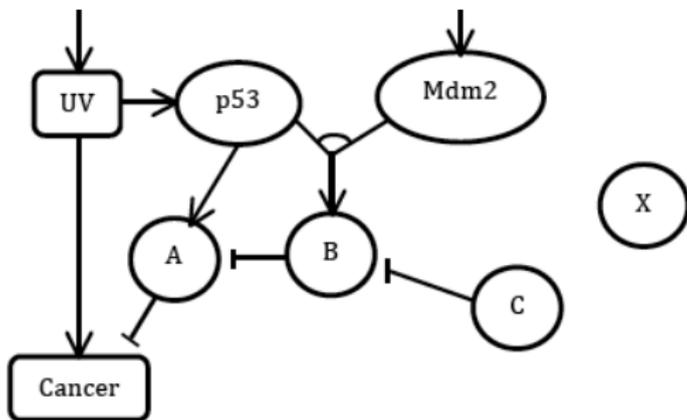
Formalisms and Algorithms based on Default Logic are proposed to study cell signaling pathways.

I - Gene Networks, Artificial Intelligence and Knowledge Representation.

The study of gene networks poses problems well identified and studied in Artificial Intelligence over the last thirty years :

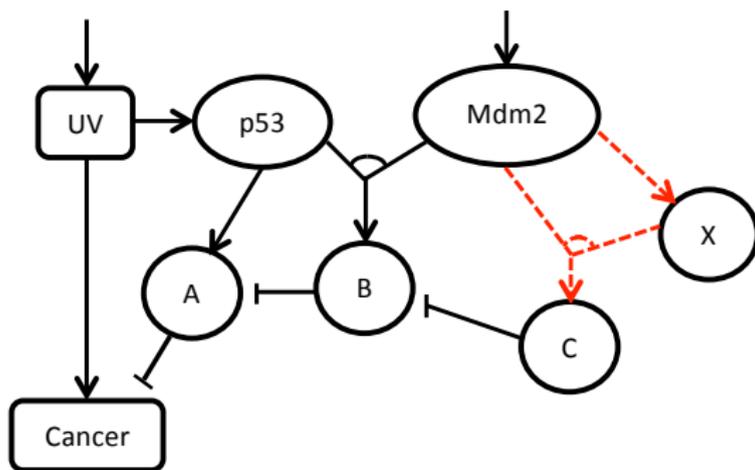
- The description of network is not complete. Biological experiments provide a number of protein interactions but not all of them.
- The conditions and sometimes the difficulties of the experiments involves these data are not always accurate.
- Some data may be very wrong and must be corrected or revised in the future.
- The information coming from different sources and experiences can be contradictory.

Different logics, and particularly non-monotonic logics, handle this kinds of situation.

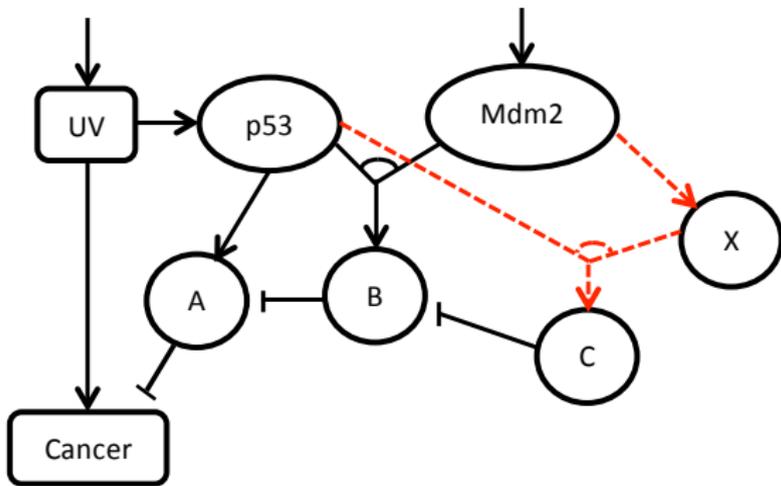


- Ultraviolet (UV) drives the cells to cancer : $UV \rightarrow Cancer$
- Protein A blocks Cancer : $A \dashv cancer$
- P53 binds Mdm2 and product B ...
 - *Triggers (UV, Cancer) and blocks (A, cancer) simultaneous !!*
 - *Contradiction - Logic Inconsistency.*

Two ways to block cancer



Add *trigger(Mdm2, X)* and *C = bind(Mdm2, X)*

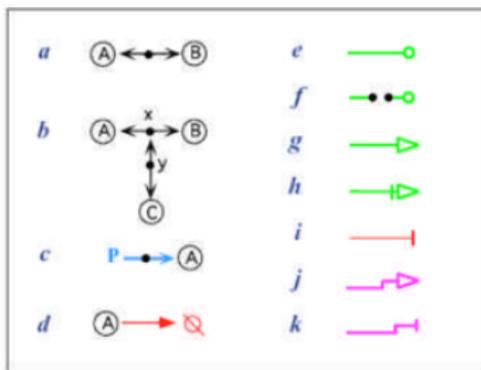


Add *trigger(P53, X)* and $C = \text{bind}(\text{Mdm2}, X)$

How to discover what to add?

ABDUCTION

Reactions



Symbols :

- (a) Proteins A and B could link, the knot represents the binding A:B ;
- (b) Multimolecular Complex binding ;
- (c) Covalent Modification of A ;
- (d) Degradation of A ;
- (e) Enzymatic stimulation in transcription ;
- (f) Autophosphorylation
- (g) General Stimulation;
- (h) Necessity ;
- (i) Inhibition ;
- (j) Activation of transcription ;
- (k) Inhibition of transcription

II - Classical Logic representation

Language representation : Propositional Logic / First Order Logic (FOL)

- Predicates :
 - $product(P)$, $trigger(P, R)$, $binding(P, Q)$, $blockBindind(P, Q)$
 - $stimulation(P)$, $phosphorylation(P)$, $dissociation..$
- Formulas / Rules :
 - Cell is subjected to ultraviolet : $stimulation(UV)$
 - Glass screen protects against UV :
 $glassScreen \rightarrow \neg stimulation(UV)$
- Concentrations :
 - $concentration(P, > 1000)$
 - $concentrationIncrease(P)$, $concentrationDiscrease(P)...$
- Time : $stimulate(dsb, dna, T) \rightarrow product(alt\text{er}ed - dna, T + 23)$
- General rules : $product(A) \wedge product(B) \rightarrow phosphorylation(A, B)$

Map of Pommier translated by 206 formulas.

Conflicts, inconsistency, causality.

Problem : Simultaneous activation and blocking :

$trigger(UV, cancer)$ and $block(A, cancer)$

- Classical Logic translations by axioms or inference rules :

(1) If A trigger B and A is true then B is true.

$trigger(A, B) \wedge A \rightarrow B$

$block(A, B) \wedge A \vdash B$

Gives inconsistency if we have informations :

$\{trigger(A, B), block(A, B), A\}$

- Simple form of causality.

Solution : Weaken the rules.

- (1-bis) If A trigger B, A is true and if **it is possible/not contradictory that B is true**, then B is true
- possible/not-contradictory \Rightarrow Non-monotonic logic. . . default logic.

III - Nonmonotonic Logics - Default Logic

- The classical logics, are monotonic :
 - If it adds information E' to a formula E , everything which was deduced from E will be deduced from $E \cup E'$.
 - Anything which is deduced from informations will always be true if we had new information.
- But in "real life" it is common to invalidate previously established conclusions when new information is added or changed.
- A non-monotonic logic allows to eliminate the monotony property of the classical logic : the conclusions could be revised with the addition of new knowledge.

- A Default Theory is a couple : $\Delta = (D, W)$
 - W set of facts (formulae)
 - D set of defaults
- A default $d = \frac{A(X) : B(X)}{C(X)}$ is a specific inference rules which handle the uncertainty.
 - If $A(X)$ is true, if it is possible that $B(X)$ is true then $C(X)$ is true.
- $A(X)$ is the prerequisite, $B(X)$ is the justification and $C(X)$ is the consequent.
- If $B(X) = C(X)$, the default is normal. A normal default means "Normally, the A are B".

EXTENSIONS .

- The goal of default logic is to find extensions of a default theory $\Delta = \{W, D\}$.
- An extension E is a consistent set of formulas obtained by adding, under condition, to W a maximal consistent set of consequents of D .
- For us, an extension can for example, represent a subgraph without conflict, of the gene network.

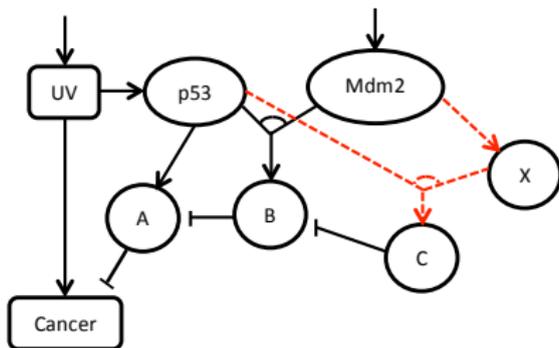
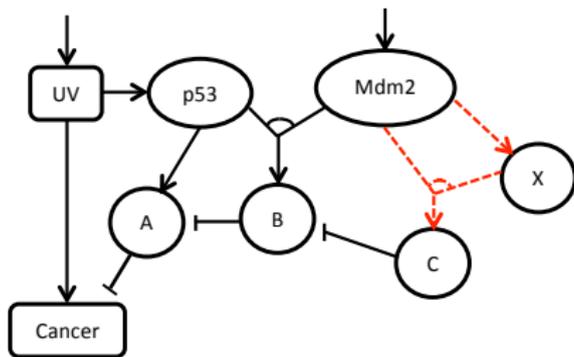
ALGORITHM (for normal defaults) :

- Starting with W and add a consequent of a default d of D .
- It is possible to use a default d if the prerequisite A is satisfied and the justification B does not lead to contradiction (the negation of B is not verified).
- If conditions are verified, add the consequent B to W and restart until all defaults has been used.

- $trigger(A, B) \wedge A \rightarrow B$ is translated by $d_1 = \frac{trigger(A, B) \wedge A : B}{B}$
- $block(A, B) \wedge A \rightarrow \neg B$ is translated by $d_2 = \frac{block(A, B) \wedge A : \neg B}{\neg B}$
- Two extensions are obtained :
 - E_1 using d_1 , which contains B
 - E_2 using d_2 , which contains $\neg B$
- Each extension provides a consistent sub-network.
- What extension prefer ?

As always other translations are possible.

IV- Abduction, discovery : Add arrows to block cancer ?



How to discover what to add ?

ABDUCTION, INDUCTION.

- Logic uses the deduction : result R is inferred from an information (formula) F

$$F \vdash R$$

- Abduction generalizes deduction : F is incomplete and abduction amounts to adding to F a consistant set of hypotheses H such that R is infered.

$$F \wedge H \vdash R$$

- Induction : information is not complete in all its generality, but we know the special cases (examples, experiences..). It should then use these cases to discover general rules.

Abduction and Induction allow to discover

Abduction by default logic

- In the case of Signaling Pathways, abduction is used mainly to search missing interactions.
- For example, the predicate $trigger(\alpha, X)$ is used inside the default with variables

$$d1 = \frac{A : trigger(\alpha, X)}{trigger(\alpha, X)}$$

$$d2 = \frac{trigger(A, B) \wedge A : B}{B}$$

α is a variable which could be instantiated by any protein.

- For the small example, by calculating the extensions which contains the predicate $block(cancer)$ we obtained 8 extensions but only 2 contains $p53$ and $mdm2$

- To represent our examples, Horn clauses are sufficient.
 - Formula $p_1 \wedge p_2 \dots \wedge p_n \rightarrow q$, with p_i and q positive literals.
 - Negative formulas : $\neg(p_1 \wedge p_2 \dots \wedge p_m)$
- SAT problem NP-complete for the general case, linear for Horn clauses. If you must leave the Horn clauses, there are strong tools to control the combinatorial explosion :
- Used for application «Simulate decisions of a commander on a submarine in wartime» - Lacroix, Siegel, Toulgoat 2011.

VI- Implementation - Language - First results.

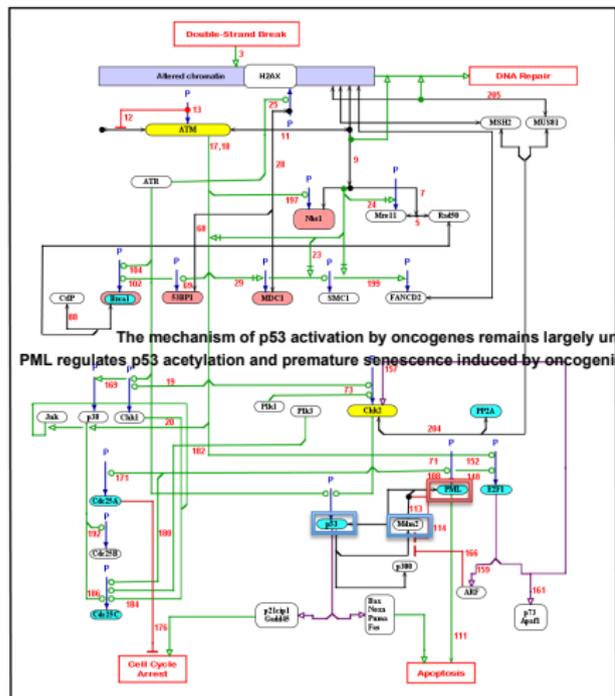
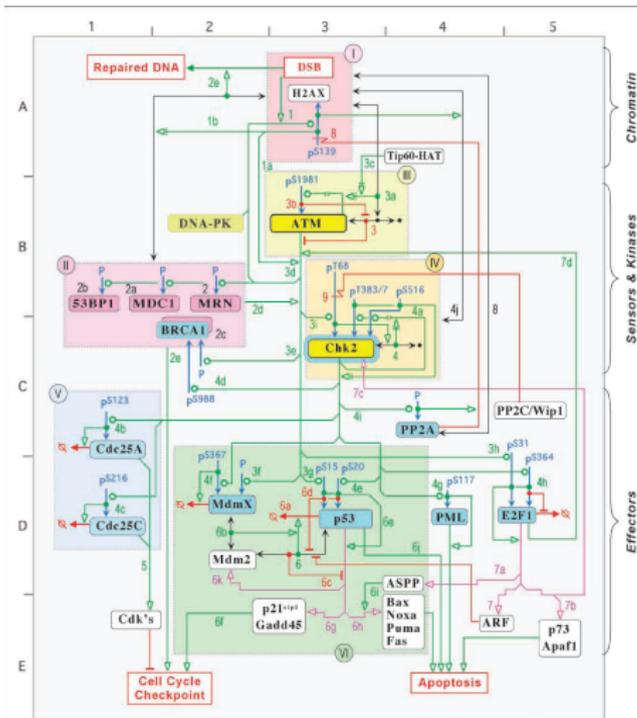
- The rules are unary clauses of Prolog.
- Map of Pommier translated very naturally by 206 rules.
- It is very simple to test the configurations by adding/removing rules and changing their nature (hard/default)
- As the clauses are Horn clauses and as the defaults are normal, the search tree can be highly optimized
- Algorithm written in SWI Prolog - Mac OS 9.

For the map of Pommier, with most defaults rules, 2 extensions calculated in 500,000 LIPS and 0.2 seconds CPU time.

A rule is a triplet ($\langle type \rangle, \langle corps \rangle, \langle weighth \rangle$).

- $\langle type \rangle$ can take 2 values : *hard* or *def*.
 - *def* : Normal default.
 - *hard* : Hard-rule, Horn clause sure and non revisable.
- $\langle weighth \rangle$ allows to weight extensions
- $\langle corps \rangle$ is a couple (L, R) .
 - L is set of literals perhaps empty, identified to $l_1 \wedge .. \wedge l_n$
 - R is either a single literal or empty.
- If the rule is hard (L, R) represent a Horn clause $L \rightarrow R$
 - Literals of L are positive, R positive or negative literal.
 - If L is empty, R is a unary clause, without variable.
 - If R is empty, L is identified to a negative clause $\neg l_1 \vee .. \vee \neg l_n$
- If the rule is $\langle def \rangle$, normal default $\frac{L : R}{R}$.
In this case L can contains negative literals.
- It is possible to introduce variables (with care).

Build-up "manually" and automatic maps.



- The AI challenge is to explain new phenomena using automatic causal discovery.
- We introduced formalism able to infer signaling pathway by using defaults approach and abductive reasoning.
- In this paper we define a new approach for the build up automatic the Double Strand Break Signaling Pathway. This map keeps only relevant proteins and it is very close to the bioregulatory network related to the histone γ -H2AX-ATM-Chk2-p53-Mdm2 pathway defined by Pommier.
- Good times computing.
- We spent a lot of time testing others formalisms and programs (Hypothesis logic, Production Fields, SOLAR, ASP..)

VII -Other tests formalisms and software.

- Hypothesis logic (Schwind-Siegel 1994 .. Doncescu-Siegel 2012).
 - Modal non-monotonic formalism bimodal simple and general.
 - Very difficult to implement and very bad algorithmic complexity.
- Using SOLAR (Doncescu - Inoue .. 2007).
 - Rules must use the negation by failure .. gazplant.
 - Limited to small examples (algorithm far too long).
- Production fields and Production Algorithm (Siegel 1988)
 - Good knowledge representation, abduction.
 - But still too high complexity
- Answer Set Programming.
 - Black Box
 - Tests inconclusive
 - Algorithms based on management loops, inefficient ?
 - A watch with a "new semantic" approach (Benhamou- Siegel 2012).

- Discrete Event Systems (not tested)
 - Very nice theoretical computer science.
 - Horrible Complexity
 - To our knowledge limited to a few genes.
- Around Learning : Abduction of the Induction
- "Classical" Approach, differential equation.
 - Tested with submarines, not very convincing.
 - Limited to small systems ?