Executable Knowledge
for rule-based modelling
of cellular signalling networks

Russ Harmer (CNRS & ENS Lyon)

+ 
Adrien Basso-Blandin (ENSL) & Walter Fontana (HMS)

+ 
John Bachman (HMS) & Pierre Boutillier (HMS) & Lucian Galescu (IHMC) & Ben Gyori (HMS)
Context
Cellular signalling

- Decentralized coordination of tissue formation and maintenance
  - *extra*-cellular ligands trigger *intra*-cellular signalling pathways to control cell growth, death, division, …
  - Perturbed in disease states, *e.g.* cancer, diabetes, …
    - *kinetic* perturbations: over-expression, knock-outs, …
    - *causal* perturbations: mutations, truncations, …
Why model signalling?

- Signalling networks, and their literature, are mind-bogglingly complicated
  - 1000s of proteins and 10s (or even 100s) of thousands of PPIs [protein-protein interactions]
  - empirical knowledge is fragmented and scattered across a vast literature
  - impossible to ‘work out’ in your head
Traditional modelling

- Modelling as a primarily ‘mental’ activity
  - identify the **key variables and their inter-dependencies** under **standard** perturbations
  - amounts to making **model-level assumptions during** model construction and ‘debugging’
  - model as ‘**synthesis** of understanding’
Traditional modelling

• To put it differently:
  
  • the **modelling** is done ‘in your head’

  • the **model** is an artifact that emerges fully-formed and is just ‘written down’

  • the model is then debugged if it fails to meet your mental ‘specification’
Modelling signalling

• Signalling needs models as ‘tools for discovery’

  • a formal reification of the modelling process [rather than just the resulting artefact] as an ‘audit trail’

  • precisely *in order to* achieve an understanding

• reverse-engineer a specification by combining empirical knowledge with (biological) inference
Modelling signalling

• Models must therefore

  • be easily extensible and modifiable [since empirical knowledge is always changing]

  • be arbitrarily perturbable [since we wish to discover, not hard-wire, their effects]

  • incorporate empirical and inferred knowledge [at various levels of detail]
Rule-based modelling

- **Formal** representation for the (10s? of 1000s of) protein-protein interactions (PPIs) in signalling
  - graph rewriting formalism
  - scalable stochastic simulation
  - pathways as causal traces
- Handles **kinetic**, but not **causal**, perturbations
Serendipity
The cognitive barrier

- Have to read many papers to find various fragments of knowledge about a single PPI

- many different ‘puzzle pieces’, at varying levels of detail, that must [somehow] be assembled into rules

- the effects of causal perturbations must be hard-wired by enumerating all cases [rather than emerging]

- not scalable for a human curator [believe me, I’ve done it]
Big Mechanism

• Seeks **causal explanations** of complex system behaviour [not ‘just’ **correlations**]

• **Machine reading** of papers, automatic **assembly** into models that yield causal **explanations** …

• The chosen **use case**: signalling pathways in cancer!
Breaching the cognitive barrier
Assembly

• Big Mechanism aims to make reading scalable and RBM provides causal explanations — once your PPIs have been formalized as rules

• The hard problem is assembly
  • combining fragments of knowledge into rules …
  • … in such a way that (apparently) conflicting information can be accommodated …
  • … and the effects of causal perturbations emerge
KAMI
knowledge aggregator & model instantiator

• Uses a **graph**-based representation of PPIs
  
  • a graph with two directed edge structures, respecting a **meta-model**:
    
    ```
    BND  
    "agent"  
    "flag"  
    MOD  
    
    "bnd?"  
    "res"  
    "reg"  
    
    BND  
    "BRK"  
    "o"  
    "o"  
    MOD  
    o  
    o  
    ```

  • uses graph rewriting to **update** and **aggregate** PPIs
KAMI

PubMed

BioPAX

&c.

Reading

Deep

import

export

KAMI

Instantiation

Annotation

RBM

&c.
KAMI
in BigM
“Grb2’s SH2-domain binds phosphorylated EGFR”

thanks to Lucian Galescu et alia!
KAMI

input

thanks to Ben Gyori
& John Bachman!
“Grb2 binds EGFR phosphorylated on Y1092”
"Grb2’s SH2-domain binds [phosphorylated] EGFR phosphorylated on Y1092"

I already know something about this interaction…

this is not yet fully automated: requires a semantic layer
“Grb2’s SH2-domain binds [phosphorylated] EGFR phosphorylated on Y1092”

this is a step of graph rewriting
“Grb2’s SH2-domain binds [phosphorylated] EGFR phosphorylated on Y1092”

this is another step of graph rewriting
“Grb2’s SH2-domain binds phosphorylated Shc”
KAMI

instantiate

Grb2(SH2e), EGFR(g,Y1092-p) -> Grb2(SH2e!1), EGFR(g!1,Y1092-p)
Grb2(SH2s), Shc(g,shc!p) -> Grb2(SH2s!1), Shc(g!1,shc!p)

independent!
These interactions use the same mechanism!

this is not yet fully automated: requires a semantic layer
KAMI
aggregate

this is a step of graph rewriting

merged (reversibly)
KAMI

**instantiate**

\[ \text{Grb2}(\text{SH2}), \text{EGFR}(g,Y1092-p) \rightarrow \text{Grb2}(!2), \text{EGFR}(g!1,Y1092-p) \]

\[ \text{Grb2}(\text{SH2}), \text{Shc}(g,shc-p) \rightarrow \text{Grb2}(!2), \text{Shc}(g!1,shc-p) \]
“Grb2-S90D does not bind EGFR”
**KAMI**

**enumeration**

only one rule for Grb2_D90

Grb2_S90(_SH2e_), EGFR(g,Y1092-p) —> Grb2_S90(_SH2!1_), EGFR(g!1,Y1092-p)

Grb2_S90(_SH2s_), Shc(g,shc-p) —> Grb2_S90(_SH2!1_), Shc(g!1,shc-p)

Grb2_D90(_SH2s_), Shc(g,shc-p) —> Grb2_D90(_SH2!1_), Shc(g!1,shc-p)

Automatic enumeration of rules
Grb2-S90D does not bind EGFR.

aa can be S or D.

This automatically propagates to the interaction with Shc too.
KAMI

equency

PubMed

Deep

Reading

KAMI

Instantiation

RBM

no rules for Grb2_D90

Grb2_S90(SH2), EGFR(g,Y1092~p) -> Grb2_S90(SH2!1), EGFR(g!1,Y1092~p)

Grb2_S90(SH2), Shc(g,shc~p) -> Grb2_S90(SH2!1), Shc(g!1,shc~p)

automated enumeration of rules
Wrapping up
Summary

• A purely formal graph rewriting foundation

  • represents knowledge and [revokable] hypotheses using formal operations of (update and) aggregation

• Model instantiation into RBM

  • automatically maintains desired [conflict] invariants and handles the effects of mutations because all enumeration is done by the machine
Automation?

• Other than an expert user, where could the steps of rewriting come from?

  • **semantics**: typically steric or functional properties of certain regions, e.g. SH2 or kinase domains
  
  • also allows for semantic checking and inference

• more general **inference** …
Work in progress

• Open re-implementation as a Python library
  • standard meta-models and meta-model transformations
  • can also be user-defined

• Based on a graph rewriting Python library
  • itself built on top of NetworkX
  • multi-level rewriting with upward propagation