Quelques approches formelles pour tester la robustesse de processus de reconstruction de rseaux

Anne Siegel

19 novembre 2012

Network inference

Goal Identify the main actors and functions involved in the response of a system.

Methods

- Data-mining. Statistics. Machine Learning...
- Metaheuristics. Search for a local optimal (genetic algorithms...).
- Optimization. Look for best-score solution (ILP).

Most approaches are discriminative:
their output is a ”single” most-probable solution.

Uncertainties appear at different stages of the identification process

- **Confidence** in the resulting predictions?
- Relevance of a **unique** solution?

Explore complete space of solutions?

Large range of inferred properties

- Topological structure. Transcriptional or metabolic network.
- Discrete dynamics. Logical rules
- Continuous dynamics. Parameter estimation.

Fluctuations in data

- Qualitative observations.
- Scoring of errors.

Is it possible to study the set of (sub)-optimal solutions?

→ **Enumeration, sampling ?**
→ **Formal methods?**

Enumeration...

Explore the space of solutions to combinatorial optimization problems which are relevant in system biology

Integer Linear Programming?

- Systems biology. Used in many frameworks (metabolism).
- Diffusion. Few software tools.
- **Expert level required**. Small modifications induce loss of efficiency.

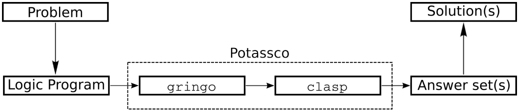
Declarative logics? (Prolog, Inductive Logic Programming...)

- Systems biology. Used mainly for experiment design.
- Diffusion. Appropriate flexibility.
- **Bad for enumeration**. Not scalable !!

→ **Find a compromise between efficiency and flexibility in the problem statement?**

Answer Set Programming: *what?* instead of *how?*

- Declarative logical problem solving paradigm
- Knowledge representation and reasoning problems
- Combinatorial search problems in NP



Potassco: **Potsdam** Answer Set Solving Collection

`http://potassco.sourceforge.net`

Rich modeling language

- Encoding problems as queries on propositional logical programs.
- *Gringo* grounder

Highly efficient inference engines

- Boolean constraint solving technology
- *Clasp* solver
- Competing with the power of SAT algorithms.

Short description

Disjunctive rules

$$\underbrace{k\,\{\,a_1;\ldots;a_n\}\,l}_{\text{head}}\ \text{:-}\ \underbrace{a_{n+1},\ldots,a_r, \mathit{not}\ a_{r+1}, ..., \mathit{not}\ a_s}_{\text{body}}$$

- Atoms. $a_1 \ldots a_n$ can be considered as facts.
- Deduction

Whenever all facts of the body are satisfied, one fact of the left part shall be true.

- Integrity constraint. ”$\leftarrow a$” is always false
- Constraint. ”a.” is always true.

Answer Set

- Set of atoms satisfying all logical rules
- Minimality and stability properties
- Every atom of an answer set appears in the head of at least one rule.

Qui a tué le docteur Lenoir ?

Program

```
3 { nom, arme, pièce } 3
1 { Colonel Moutarde, Mademoiselle Rose } 1 :- nom
1 { chandelier, revolver } 1 :- arme
1 { cuisine, hall, salon, salle à manger } 1 :- pièce
1 { cuisine, hall, salon } 1 :- Colonel Moutarde
Salon :- Colonel Moutarde, not revolver
:- cuisine
Chandelier
```

Answer Sets??

Qui a tué le docteur Lenoir ?

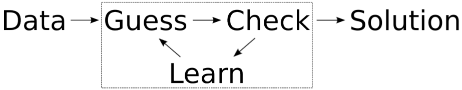
Program

```
3 { nom, arme, pièce } 3
1 { Colonel Moutarde, Mademoiselle Rose } 1 :- nom
1 { chandelier, revolver } 1 :- arme
1 { cuisine, hall, salon, salle à manger } 1 :- pièce
1 { cuisine, hall, salon } 1 :- Colonel Moutarde
Salon :- Colonel Moutarde, not revolver
:- cuisine
Chandelier
```

Answer Sets??

- Colonel Moutarde, salon, chandelier
- Mademoiselle Rose, salle à manger, chandelier
- Mademoiselle Rose, salon, chandelier
- Mademoiselle Rose, hall, chandelier

Guess & Check methodology



Data → Guess → Check → Solution

Learn

- Data: PKN and phospho-proteomics dataset (facts)

```
node(tnfa). node(p38). edge(tnfa,p38,1). exp(1,tnfa,1). obs(1,p38,0).
```

- Guess: Generate candidates models (non-deterministic)

```
{clause(A,N)} :- hyperedge(A,N).
```

- Check: Eliminate invalid models (integrity constraints)

```
:- clause(A,N), clause(B,M), A!=B, redundant(A,B).
```

- Learn: Loop between ”guess” and ”check”
- Optimize: Minimize cost function (weighted sum of atoms)

```
#minimize[mismatch(E,R,W) = W, clause(A,N) : param(P) = N*P].
```

Ongoing issue

ASP technologies are now proved to be mature and very efficient in several computational issues.
→ *constraint satisfaction, diagnosis, repairing, planning...*

Is ASP useful in systems biology?

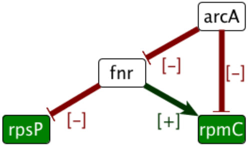
Work in progress...

- Consistency checking of network
- Inference of logical rules for signaling networks
- Inference of robust regulatory nodes
- inference of metabolic network

Validation/Correction of (possibly infered) networks

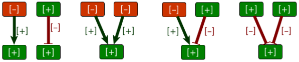
Knowledge-representation

- Regulations. Signed oriented graph.
- Edge colors. Regulatory effects.
- Node colors. Expression data.



Constraint over graph-coloring

- Causal law. Explain the expression of each target gene by the consistent regulation of a source
- Forbidden patterns.



FORBIDDEN COLORED PATTERNS

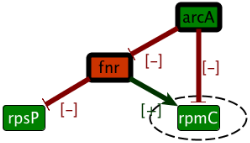
ASP encoding

```
vertex(rpsP). vertex(fnr).
vertex (arcA). vertex(rpmC).
edge(fnr,rpsP). observedE(fnr,rpsP,-).
edge(fnr,rpmC). observedE(fnr,rpmC,+).
edge(arcA,fnr).  observedE(arcA,fnr,-).
edge(arcA,rpmC). observedE(arcA,rpmC,-).
observedV(rpsP,-).  observedV(rpmC,-).
```

```
labelV(I ,+) ; labelV (I ,-) ← vertex(I).
labelV(I ,S) ← observedV(I,S).
labelE(J,I,+) ; labelE (J,I,-) ← edge(J,I).
labelE(J,I,S) ← observedE(J,I,S).
receive(I,+) ← labelE(J,I,S), labelV(J,S).
receive(I,-) ← labelE(J,I,S), labelV(J,T), S≠T.
← labelV (I,S), not receive(I,S).
```

Results *[Guziolowski-BMCGenomics'09, Gebser-KR'10]*

- Prediction. *rpsP* and *fnr* have fixed colors according to allowed patterns.
- Diagnosis. An extra forbidden pattern appears on *rpmC*.
- Correction. Also possible.

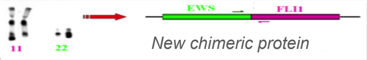


The expression of rpmC cannot be explained.

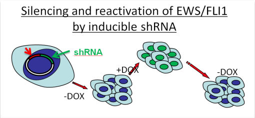
Example of application

Ewing Sarcoma

- Chimeric protein
- Institut Curie. Inactivation of the protein expression.



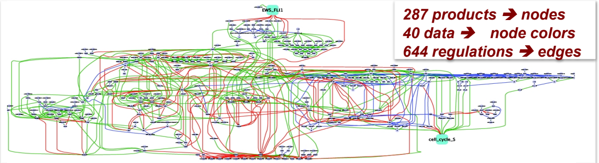
New chimeric protein



Silencing and reactivation of EWS/FLI1 by inducible shRNA

Data *[Institut curie. Barillot & Delattre]*

- Litterature-based regulatory network
- Time-series genes expression after the protein inactivation



287 products → nodes

40 data → node colors

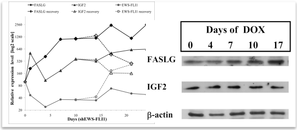
644 regulations → edges

→ What can be surely predicted from this information?

Cancer application

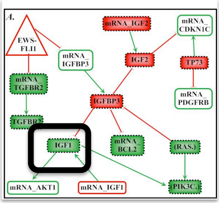
Explain and predict

- Effect of multi-scale competitions.
- Validation of predictions.



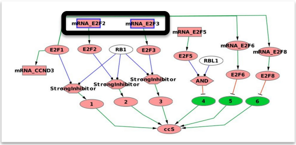
Key pathways *[Baumuratova-BMC syst. bio'10]*

- Missing regulations over IGF1



Design? *[Guziolowski TCBB'11]*

- Two new possible targets for EWI-FLI1
- si-RNA confirmation (unpublished)

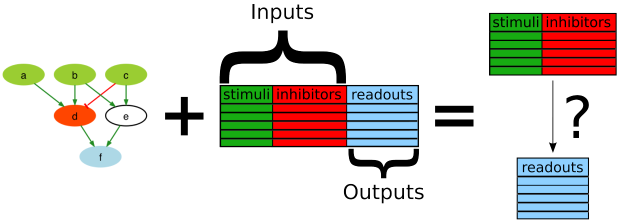


Learning logical static rules

Data

- Signed and directed causal interactions among proteins
- Phosphorylation activity in time *t* after stimulation

Goal Predictive models of immediate-early protein signaling pathways



Prior Knowledge Network

Phospho-proteomics Dataset

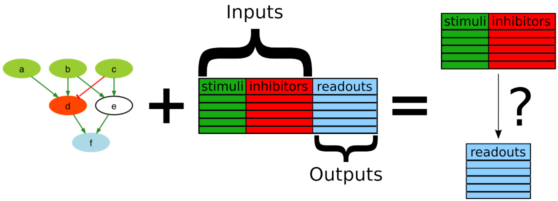
Predictive Model

Underlying assumptions

- Focus on fast reactions
- No time for feedback mechanisms
- Pseudo-steady state assumption

Predictive Signaling Network Challenge *[Prill'11]*.

12 groups with different formalisms (ODEs, machine learning, boolean logic)



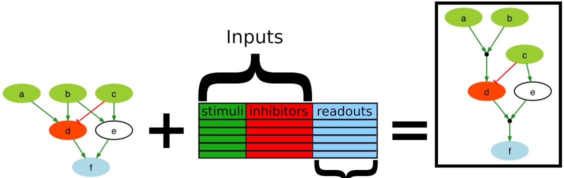
Prior Knowledge Network Phospho-proteomics Dataset Predictive Model

Score **Trade-off between fitness and model size**.

- Biological Property: consistency with experimental data
- Parsimony Principle: minimal/simplest explanation

Discrete approach

Optimization Learning **Logic** Models or **hypergraphs**?



$(a \wedge b) \vee \neg c \iff d$

$c \iff e$

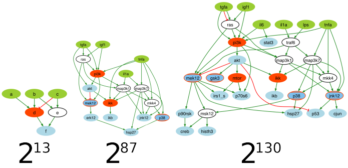
$d \wedge e \iff f$

Prior Knowledge Network

Phospho-proteomics Dataset

Optimal Logic Model (Hypergraph)

Search space. Hypergraphs compatible with the graph ($2^{\#hyperedges}$)

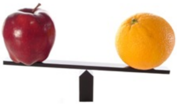


2^{13} 2^{87} 2^{130}

State-of-the-Art CellNOpt *[Saez-Rodriguez’09]*

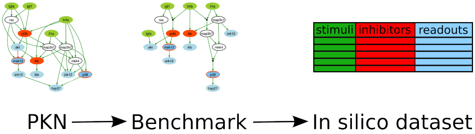
- Genetic algorithm to train logic models
- Weaknesses **Guarantee to find all global optimal models? Scaling?**

Comparing meta-heuristics and declarative logics



Benchmark and comparison sets

- 2 real cases. Middle and large-scale with discretized real dataset.
- 240 in-silico cases. middle-scale, several benchmarks and in-silico datasets.



stimuli	inhibitors	readouts

PKN → Benchmark → In silico dataset

Criteria of comparison

- Success / Completeness
- Time performance

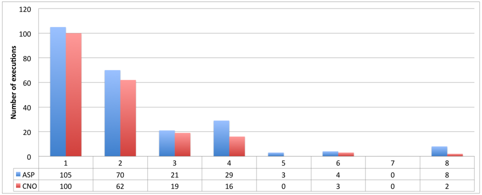
Success / Completeness

Discretized real datasets *[Videla-CMSB'12]*

Optimal models	ASP	CellNOpt
Middle	8	2
Large	2	**0**

- Several optimal models.
- Metaheuristic miss all optimals in the large-scale case.

Generalization: 240 in-silico studies *[Videla-CMSB'12]*



	1	2	3	4	5	6	7	8
🟦 ASP	105	70	21	29	3	4	0	8
🟥 CNO	100	62	19	16	0	3	0	2

- **No single optimal model in more than 60% of studies**.
- Metaheuristics fail in identifying all optimal when they are numerous.

Time performance

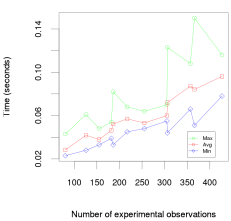
Discretized real datasets

- **Metaheuristics are perturbed by the identification of global optimal models**

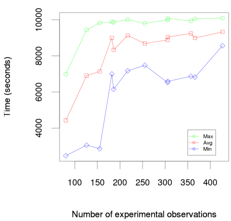
Times	ASP (s)	CellNOpt (h)
Middle	0.09 seconds	9.2 hours
Large	0.5 seconds	27.8 hours

Generalization: 240 in-silico studies

ASP computation time



CellNOpt computation time



- Significant improvement in computation times

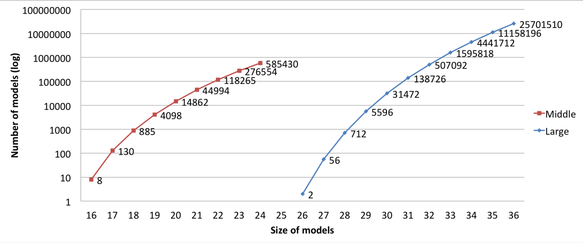
Analysis of incompleteness

Metaheuristics

	Number of saved models	Size	Minimal size
Middle	66	16→24	16
Large	206	27→36	26

- CNO finds suboptimal models “close” to optimal models
  → Are they a good representation of the space of sub-optimal models?

Space of sub-optimal models?



- **ASP allows enumerating** the space of suboptimal models
- Exponential growth with the size
- *No information on the representativity of CNO models*

Impact of real data? Space of sub-optimal models?

Real case *[Videla-Work in progress]*

- Numerical value → 100-value discretizations !
- Loops → new encoding
- Good time performances

(Non)-unicity of solutions

- Still several optimal solutions
- Combinatorics: **mutual exclusive patterns**.

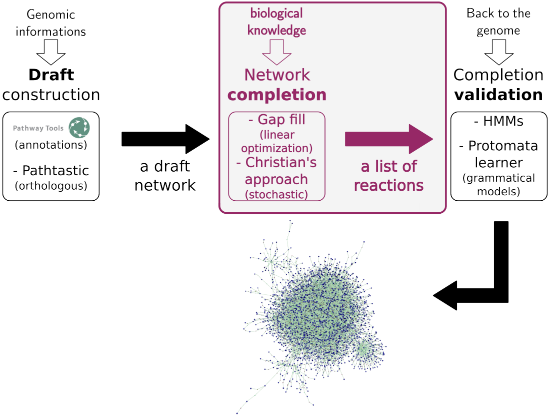
Including noise? *[Guziolowski-Work in progress]*

- A 10% noise over real data is inherent to the technologies.
- **Enumeration: more than 10000 sub-optimal models**

→ **Relevance? Strong need for biological metrics to select models!**

Quite troubling...

Optimization methods → Metabolic network completions



Genomic informations

Draft construction

Pathway Tools (annotations)

- Pathtastic (orthologous)

a draft network

biological knowledge

Network **completion**

- Gap fill (linear optimization)
- Christian's approach (stochastic)

a list of reactions

Back to the genome

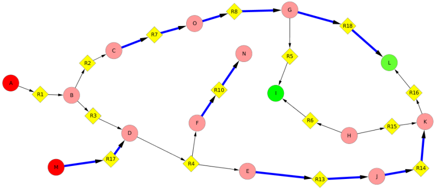
Completion **validation**

- HMMs
- Protomata learner (grammatical models)

Hints on the number of possible completions?

Optimization-based methods

Cardinal minimal completion Add the minimal number of reactions to explain the presence of metabolites.



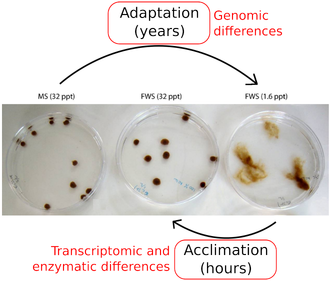
Linear programming: GapFill *[Kumar'07]*

- Single-based completion. *For each compound*, add the minimal number of reactions in the network

ASP *[Thiele'11]*

- Global-based completion. Add the global minimum of reactions in the network required *for all compounds in the same time*

Application to an eukaryot example



Adaptation (years)

Genomic differences

Transcriptomic and enzymatic differences

Acclimation (hours)

Brown algae

- Ectocarpus. Model for brown macro-algae.
- Specifities. Very distant from well-studied green micro-algae.
- Capable of adaptation and acclimation.

Data *[Station Biologique Roscoff]*

- Genome. High-quality annotated genome.
- Metabolism. List of 56 characterized metabolic compounds.

Ectocarpus metabolism reconstruction

Reconstruction *[Prigent. Work in progress]*

- Automatic tools. Bad reconstruction.
- Global cardinal completion. **Adding** 59 **metabolic reactions allows producing 48 compounds over 56**
- Single-based completion. 38 reactions belong to all solutions.

Enumeration *[Prigent&Thiele. Work in progress]*

- ASP. Enumeration is possible.
- Combinatorial explosion. **The full set of possible completions contains 16 millions of solutions**
- Reactions. About 100 reactions occur in at least one solution.
- Performance. High level of RAM. Extreme range of solvers.

Current issues

- New biological metrics to sort information !
- Integration Take advantage of the flexibility of the declarative language to insert new critera of classification.
- Sampling the space of solution? [*Christian'11]*

Conclusion

Novelties brought declarative logic paradigms ??

High-level declarative language

→ Easy ”step-by-step” encoding of data integration and constraints.

- Confrontation of a reconstructed network with additional data.
- Learn the logic quasi-steady state response of signaling networks.
- Completion of metabolic networks.

Enumeration of complete space of solutions

→ Explore the combinatorics responsible of the explosion of the size

- Global correction of transcriptional networks.
- Sub-optimal solutions to middle-case problems (learning the dynamics).
- Global set completion to metabolism reconstruction.

Work in Progress

- Biological criteria to elucidate the structure of the space of solutions.
- Sampling issues.
- ASP: backtrack traces of ”proofs” to classify the importance of initial information.
- Software.

Acknowledgements

ASP Modelling

- IRISA/Inria (Rennes). S. Videla (PhD-ANR Biotempo). S. Prigent (Ph-D) . S. Thiele (Post-doc).
- LINA (Nantes). D. Eveillard (MCF), J. Bourdon (MCF).
- Ircyyn (Nantes). C. Guziolowski (MCF)
- Univ. Luxembourg. T. Baumuratova (Post-doc).

ASP solvers and grounders

- Potsdam university. T. Schaub (Prof). M. Gesber (Ass. Prof). M. Ostrowski (PhD).
- Rennes. J. Nicolas (DR Inria).

Biological applications

- Institut Curie (Ewing sarcoma). G. Stoll. D. Surdez. O. Delattre. A. Zinoviev. E. Barillot.
- EBI (signaling network). J. Saez-Rodriguez. F. Eduarti.
- Station Biologique Roscoff (Algae). T. Tonon. C. Boyen.