Design using intuition and logic

Optimisation, Graphical Models, Protein Design

T. Schiex, INRA MIAT

S. de Givry, G. Katsirelos, D. Simoncini (IRIT), S. Barbe (TBI/INSA)

ANITI September 2019

Context: progress in NP-complete Problem Solving

SAT

- Canonical NP-complete problem (Cook theorem)
- A set X of Boolean variables
- A set C of clauses (disjunction of litterals: a variable or its negation)
- $\exists$? a labelling of X such that all C is true

SAT solvers find a solution or provide a proof that none exists

- Major impact on digital circuit verification (PSPACE-complete),...
- Theorem proving (recent proof on Pythagorean Triangles[9,10])
- Millions of variables, 10s of millions of clauses

Main ingredients

A lot of empirical work

- Lots of real problems (random problems are different)
- Competitions with Open Source software

Main elected ingredients

- Massive problem reformulation using local inference (Unit Propagation, fast data-structures)
- If insufficient, make assumptions (tree search)
- Make non naive assumptions (adaptative variable ordering, learned during search)
- Conflict analysis (clause learning following inconsistent assumptions)
- Restarts,...

Constraint network (X, C) — Joint feasibility distribution

- a sequence X of variables x_i, finite domain D_i
- a set C of constraints — boolean functions
- $c_S \in C$ involves variables in $S \subseteq X$ — table, clause,...

$$\prod_{i \in S} D_i \rightarrow \{t, f\}$$

- Joint boolean function $F(X) = \bigwedge c_S$

Applications

Scheduling, rostering, planning, configuration...

SAT and CSP

Excellent to describe, analyze, design perfectly known complex systems

SAT and CSP

Excellent to describe, analyze, design perfectly known complex systems

Biology/Life

Full of imperfectly known complex systems

From boolean to numerical functions

Cost function network (X, W) — Joint cost/feasibility distribution[3,15]

- a sequence X of variables x_i, finite domain D_i
- a set W of cost functions — $w_\varnothing$ (lb)
- $w_S \in W$ — table/tensor, clause, simple function...

$$\prod_{i \in S} D_i \to \{0, \ldots, k\}$$

- Joint cost function $W(X) = \sum w_S$ — (bounded sum)

Central problems: WCSP (Partial Weighted MaxSAT)

- solution: cost less than k
- optimal: w.r.t. the joint cost $W(X)$ — decision NP-complete
- constraint: function with costs in $\{0, k\}$ — CP is $k = 1$

Graphical models (originally a stat/ML terminology)

GMs define

- a joint function of many variables
- by combining (using a dedicated operator)
- a set of simpler functions (scopes, langage)

What function, what query?

- feasibility: prop. logic, constraint nets (CSP: $\vee, \wedge$)
- priorities: possibilistic/fuzzy CSP $(\max, \min)$
- cost, energy: Cost Function Networks (WCSP: $\min, +$)
- probability: Markov Random Field, Bayes nets (Max. a posteriori: $\max, \times$) (Marginal: $+, \times$)

Extended most ingredients from SAT/CSP solvers

- Incremental reformulation techniques (tighter lower bound)[4]
- Making assumptions (Hybrid Branch and bound, lb. $w_\varnothing$)
- Non naive variable ordering (adaptative)
- Graph decomposition (treewidth combined with all the above)
- Dominance analysis (Dead End Elimination)
- Still missing: conflict analysis

Open source Toulbar2 solver

- Won several competitions (on approximate MAP/MRF solving)
- "ToulBar2 variants were superior to CPLEX variants in all our tests"[7]

Applications

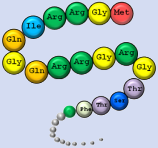
- Life sciences: protein design, genotyping data diagnosis and repair, RNA gene finding, crop allocation
- NLP, music composition (MLN), Data mining, timetabling, planning, POMDP, universal Hashing based counting, probabilistic inference, Inductive LP, image processing...
- see toulbar2 web site and GitHub

Proteins

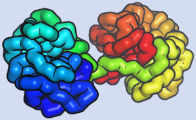
Most active molecules of life

Sequence of amino acids, 20 natural ones each defined by a specific flexible side-chain

Folding



$\longrightarrow$



$\longrightarrow$ Function

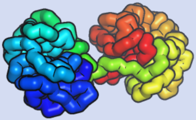
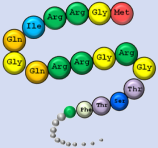
Transporter, binder/regulator, motor, catalyst...
Hemoglobine, TAL effector, ATPase, dehydrogenases...

Protein Design

Most active molecules of life

Sequence of amino acids, 20 natural ones each defined by a specific flexible side-chain

Inverse folding

Function →  → 

Transporter, binder/regulator, motor, catalyst...
Hemoglobine, TAL effector, ATPase, dehydrogenases...

Why is it worth designing new proteins?

ANITI

Eco-friendly chemical/structural nano-agents

- Biodegradable (have been mass produced for billions of year)
- "Easy" to produce (transformed bacteria)
- Useful for health, green chemistry[14] (biœnergies), nanotechnologies[17]...

20^n sequences! experimentally intractable

Energy optimisation side - NP-complete

- efficient exact energy optimisation for protein design (far faster than ILP,[1] compete with simulated or D-Wave quantum annealing[11,16])
- specific extensions for Protein Design: counting, multi-state (flexibility)

Actual protein designs

-  A self assembling hyper-stable protein[17] (with A. Vœt, KU Leuven)
- New light-weight antibody with nice properties (with A. Olichon, Toulouse Cancer Research Center)

Logical and probabilistic propositional reasoning

- satisfy logical properties/constraints exactly
- optimise a criteria that can be probabilistic (or not)
- which can be learned from data (likelihood/convex optim.).

Protein Design

- desired design properties (logical information),
- physical knowledge (represented as a decomposable energy function)
- probabilistic information learned from data (known protein sequences)

-joint project: guaranteed relational probabilistic/logic reasoning

Build a rigorous platform (Markov Logic Networks,[13] Soft Probabilistic Logic,[2] ProbLog[5])

Topics for ANITI

Topics

- stronger lower bounds: convex/SDP relaxations. Λ-PhD.
- learn when to use them, better heuristics (Multi-Armed Bandits, NN).
- extend conflict analysis to CFNs (through duality). Λ-PhD
- learn CFNs (available for numerical information)
- parallelization, CPD application, PhDs : Λ-PostDoc
- Consider multiple protein geometries: Quantified WCSP (bi-level optimisation). ANR SPaceHex.

ANITI links

Reasoning with rules and data

- Useful for other chairs? (argumentation, NLP, ...)
- Renault and configuration: learning from history (fairness/biases)
- Learning optimally sparse and proving properties of ML models[8,12]
- DL for CPD (adversarial, transformer).

Continuous optimisation

- Fast incremental convex lower bounds
- Continuous movements: non convex hybrid (discrete/continuous) optimisation problem [6]
- Tight link with robotics (side-chains are robotic arms, J. Cortes/LAAS/CNRS. PhD).

ANITI

[1] David Allouche et al. "Computational protein design as an optimization problem". In: *Artificial Intelligence* 212 (2014), pp. 59–79.

[2] Stephen H Bach et al. "Hinge-loss markov random fields and probabilistic soft logic". In: *arXiv preprint arXiv:1505.04406, JMLR* (2015).

[3] M. Cooper et al. "Soft arc consistency revisited". In: *Artificial Intelligence* 174 (2010), pp. 449–478.

[4] Martin C Cooper et al. "Soft arc consistency revisited". In: *Artificial Intelligence* 174.7 (2010), pp. 449–478.

[5] Luc De Rædt, Angelika Kimmig, and Hannu Toivonen. "ProbLog: A Probabilistic Prolog and Its Application in Link Discovery.". In: *IJCAI*. Vol. 7. Hyderabad. 2007, pp. 2462–2467.

[6] Abram L Friesen and Pedro Domingos. "Recursive decomposition for nonconvex optimization". In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*. 2015.

[7] Stefan Haller, Paul Swoboda, and Bogdan Savchynskyy. "Exact MAP-Inference by Confining Combinatorial Search with LP Relaxation". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[8] Guy Katz et al. "Reluplex: An efficient SMT solver for verifying deep neural networks". In: *International Conference on Computer Aided Verification*. Springer. 2017, pp. 97–117.

[9] Oliver Kullmann. "The Science of Brute Force". In: *Communications of the ACM* (2017).

References II

[10] Evelyn Lamb. "Maths proof smashes size record: supercomputer produces a 200-terabyte proof–but is it really mathematics?" In: *Nature* 534.7605 (2016), pp. 17–19.

[11] Vikram Khipple Mulligan et al. "Designing Peptides on a Quantum Computer". In: (2019).

[12] Nina Narodytska et al. "Verifying properties of binarized deep neural networks". In: *Proc. of AAAI'18*. 2018.

[13] Matthew Richardson and Pedro Domingos. "Markov logic networks". In: *Machine learning* 62.1-2 (2006), pp. 107–136.

[14] Daniela Röthlisberger et al. "Kemp elimination catalysts by computational enzyme design". In: *Nature* 453.7192 (2008), p. 190.

[15] T. Schiex, H. Fargier, and G. Verfaillie. "Valued Constraint Satisfaction Problems: hard and easy problems". In: *Proc. of the 14th IJCAI*. Montréal, Canada, Aug. 1995, pp. 631–637.

[16] David Simoncini et al. "Guaranteed discrete energy optimization on large protein design problems". In: *Journal of chemical theory and computation* 11.12 (2015), pp. 5980–5989.

[17] Arnout RD Vœt et al. "Computational design of a self-assembling symmetrical β-propeller protein". In: *Proceedings of the National Academy of Sciences* 111.42 (2014), pp. 15102–15107.