Star-AI for the Analysis of Gene Data

Ralf Möller

Presentation is based on:

Nikita, Sakhanenko, David Galas.  
Markov Logic Networks in the Analysis of Genetic Data  
Knowledge-based Genotype-Phenotype Associations

- Genome-wide association studies (GWAS) and similar statistical studies of g-p-linkage data assume **simple additive models of gene interactions**
  - Viewing compound effects of multiple genes on a phenotype as a **sum of influences** of each gene often misses a substantial part
  - Methods **do not use** any biological knowledge about underlying mechanisms
  - Unconstrained association searches require too many population samples, and can succeed only in detecting a limited range of effects

- **Goal:** Incorporate biological knowledge into statistical analysis
  - Need **probability theory** to capture uncertainty
  - Need FO Logic to avoid “model explosion”

- **Claim:** Can use Stochastic Relational AI (Star-AI) as an enabler
  - Complex, non-additive genetic interactions modeled
  - Learning with datasets of “reasonable” size
Model Joint Distribution w/ Markov Random Field

MRFs are too simple:
Model “explodes” with more complex interaction on many genes?

Propositional Logic
Why don’t we use First-Order Logic?
FO-Signature for Biological Knowledge Base

- **RelWS(x, v)**
  - 2-argument predicate which captures a relation between a *wild* type and a *single* mutant
- **RelWD(x, y, v)**
  - Relation between a *wild* type and a *double* mutant
- **RelSS(x, y, v)**
  - Relation between two *single* mutants
- **RelSD(x, y, x, v)**
  - Relation between a *single* mutant and a *double* mutant
- **Int(x, y, c)**
  - Interaction between two genes with interaction type c

- **x, y**: Genes: Domains \{g1, g2\}
- **v**: Phenotype values: Domain \{0, 1, 2\}
- **c**: Interaction types: Domain \{A, B\}
Dependencies

Just a subgraph
Objects, Relations, Uncertainty about Formulas

Markov Logic Network (MLN)

• Depending on the type of interaction between two genes \( \text{Int}(x, y, c) \), there is a dependency between \( \text{RelWS}(x, v) \) and \( \text{RelSD}(y, x, y, v) \)

• Three statements \( \text{RelWS}(x,v) \), \( \text{RelWS}(y,u) \), and \( \text{RelWD}(x,y,w) \), together determine the type of gene interaction \( \text{Int}(x, y, c) \)

1.5 \[ \forall x, y \in \{g1, g2\}, \forall c \in \{A, B\}, \forall v, u \in \{0, 1, 2\}, \]
\[ \text{Int}(x, y, c) \Rightarrow (\text{RelWS}(x, v) \Leftrightarrow \text{RelSD}(y, x, y, u)) \]

2.1 \[ \forall x, y \in \{g1, g2\}, \forall c \in \{A, B\}, \forall v, u, w \in \{0, 1, 2\}, \]
\[ \text{RelWS}(x, v) \land \text{RelWS}(y, u) \land \text{RelWD}(x, y, w) \Rightarrow \text{Int}(x, y, c). \]

Mathematics behind this representation language:

• Joint probability distribution over ground atoms (in this case Boolean randvars)
• Definition of jpd based on weights associated with formulas (details omitted for brevity)
Query Language

• What is the probability that a ground atom $Q$ is true (event) given that every ground atom from a set $\{E_1, \ldots, E_m\}$ is true (conjunction of evidences)?

$$\Pr(Q \mid E_1 \land \ldots \land E_m, MLN)$$

• Query semantics based on MLN groundings

$$\Pr(Q \mid E_1 \land \ldots \land E_m, MLN) = \frac{\Pr(Q \land E_1 \land \ldots \land E_m \mid MLN)}{\Pr(E_1 \land \ldots \land E_m \mid MLN)} = \frac{\sum_{\gamma \in \Gamma_P \cap \Gamma_E} \Pr(\gamma \mid MLN)}{\sum_{\gamma \in \Gamma_E} \Pr(\gamma \mid MLN)},$$

where $\Gamma_P$ is the set of all possible configurations where a ground predicate $P$ is true, and $\Gamma_E = \Gamma_{E_1} \cap \ldots \cap \Gamma_{E_m}$. 
Application: Yeast Sporulation

• Set of 374 progeny of a cross between two yeast strains (a wine and an oak strain) differing widely in their efficiency of sporulation

• For each of the progeny, the sporulation efficiency (phenotype) was measured and assigned a value from {very_low, low, medium, high, very_high}

• Each yeast progeny strain was genotyped at 225 markers uniformly distributed along the genome
  • Each marker takes on one of two possible values indicating whether it derived from the oak or wine parent genotype

Example Knowledge Base and its Use

Model the effect of a single marker on the phenotype, i.e., sporulation efficiency:

- **Signature of the model**
  - $G(s, m, g)$: Markers’ genotype values across yeast crosses (evidence, predictor)
  - $E(s, v)$: Phenotype (sporulation efficiency) across yeast crosses (target)
  - $s$: Strain
  - $m$: Marker
  - $g$: Genotype value (wine or oak parent)
  - $v$: Phenotype value (very_low, ..., very_high)
  - $w$: Weight of the formula

- **Information need**: Find optimal strains

- **KB**: MLN patterns: $\forall \text{strain} \in \{1, ..., 374\}, G(\text{strain}, m, g) \Rightarrow E(\text{strain}, v), w_m, g, v,$

- **Semantics**: Formulas and their weights define probability distribution

- **Queries**: $P( E(\text{Strain, very_high})=true \mid G(42, m_1, g_1)=true, ..., G(42, m_{17}, g_{23})=true )$

- **Answer to satisfy information need**: Return strains with k-highest probability values
MLN Query Answering Algorithms

• Naïve grounding (combinatorial), then ground MRF QA
• Clever grounding (consider only relevant groundings, still combinatorial)
• Sampling (maybe quite inexact, approximation quality hard to control)
• Lifted query answering (exact, FPT: exponential in “tree width”, which is fixed for a model and small, linear in size of variable domains for liftable query classes)
  • Our preparatory works:

  Tanya Braun. 
  Rescued from a Sea of Queries: Exact Inference in Probabilistic Relational Models
  Dissertation 2020

  Tanya Braun, Ralf Möller, Marcel Gehrke.
  https://www.ifis.uni-luebeck.de/index.php?id=672
  Tutorial at ECAI 2020
MLN Learning from Application Data

Estimate ground joint probability distribution from data

Learning goal: Encode jpd in sparse form using MLNs

• Full MLN learning:
  • Take model signature from database schema
  • Determine suitable formulas from predicates in signature
  • Determine weights using maximum likelihood estimator

• Weight learning only (formulas given):
  • Determine weights using maximum likelihood estimator

Lise Getoor, Ben Taskar. 
Introduction to Statistical Relational Learning. 
Challenges for Research

• Develop Intelligent Agents for
  • Finding optimal targets for given predictors: Apply approach to gene analysis problems
  • Allow for cooperating agents to organize learning autonomously
    • Generalize results from precision medicine

• Deal with interaction of gene sequences in a genome rather than single genes?
  • Exploit results on temporal reasoning (dynamic Star-AI)?
  • Our preparatory work:

  Marcel Gehrke, Ralf Möller, Tanya Braun.
  Taming Reasoning in Temporal Probabilistic Relational Models

  Marcel Gehrke.
  Taming Reasoning in Temporal Probabilistic Relational Models
  Dissertation 2021

• Compile MLNs into Lifted Tensor Networks for faster execution on a quantum computer?
  • Exploit entanglement of qubits in a lifted way to compute with “reasonable” number of qubits

  Nathan A. McMahon, Sukhbinder Singh & Gavin K. Brennen.
  A holographic duality from lifted tensor networks.
  npj Quantum Information volume 6, Article number: 36. 2020.
Bibliography

Application scenario:


See also:


For QA as well as learning algorithms for Star-AI, see:

• https://www.ifis.uni-luebeck.de/index.php?id=672

• https://www.ifis.uni-luebeck.de/index.php?id=703&L=2