# Reasoning with imprecise probabilistic knowledge on enzymes for rapid screening of potential substrates or inhibitor structures

Weiru Liu[1] Anbu Yue[1] David J Timson[2]

School of Electronics, Electrical Engineering and Computer Science,
Queen's University Belfast, Belfast BT7 1NN, UK
{w.liu, a.yue}@qub.ac.uk

School of Biological Science
Queen's University Belfast, Belfast BT9 7BL, UK
d.timson@qub.ac.uk

July 18, 2009

# Outline

# Introduction

- Modeling imprecise Knowledge.
- Analyze how useful a knowledge base is for answering queries.
- Extract reliable and meaningful answers.
- A system for querying, analyzing, revising, and merging imprecise probabilistic knowledge bases.
- Apply this system for substrate prediction.

# Why Conditional Probabilistic Logic Programming?

- There have been a lot of research efforts directly towards integrating logical programming with probability theory.
  - Causal Probabilistic Logic Programming [C. Baral, M. Gelfond, and J. Rushton, LPNMR2004].
  - Success Probabilistic Logic Programming[N. Fuhr, JASIS, 2000]
  - The work proposed in [A. Dekhtyar and M. Dekhtyar, ICLP2004]
  - etc..
- Our project focus on modeling and handling (inferring, revising, merging, etc.) clinical trials.
- Conditional probability is more suitable to model statistical data from clinical trials.
  For example, $(mortality|treatment\_name, disease\_name)[l, u]$
- Conditional Probabilistic Logic Programming can be implemented [T. Lukasiewicz, ACM Trans. Comput. Log., 2001]

# Preliminary: conditional Probabilistic Logic Program (PLP)

- An *event* or *formula*, denoted as $\phi, \psi$, *etc*., is of the form $p(X_1, \ldots, X_m), q(t_1, \ldots, t_n)$, *etc*.
- A *probabilistic formula* is of the form $(\psi|\phi)[l, u]$
- $I \models_{cl} \phi$, iff $I \models_\sigma \phi$ for all assignment $\sigma$.
- A *probabilistic interpretation Pr* is a probability distribution on all possible worlds.
- $Pr_\sigma(\phi) = \sum_{I \in \mathcal{I}_\Phi, I \models_\sigma \phi} Pr(I)$.
- $Pr \models_\sigma (\psi|\phi)[l, u]$ iff $Pr_\sigma(\phi) = 0$ or $Pr_\sigma(\psi|\phi) \in [l, u]$.
- $Pr \models (\psi|\phi)[l, u]$ iff $Pr \models_\sigma (\psi|\phi)[l, u]$ for all assignment $\sigma$
- Let $P$ be a PLP, $Pr \models P$ iff $Pr \models \mu$ for all $\mu \in P$.
- Let $P$ be a PLP, $P \models (\psi|\phi)[l, u]$ iff for all $Pr \models P$, $Pr \models (\psi|\phi)[l, u]$

# Measuring the ignorance and degree of satisfaction

- **Ignorance** The measure of ignorance (w.r.t. a query) reflects how reliable a precise probability for query can be. A high value of ignorance suggests that a single probability is not suitable for answering the query, and we can only expect a non-informative interval.
- **Degree of satisfaction** Measure the (second order) probability that the exact probability of a query falls in a user-given interval. One bound with sufficient degree of satisfaction gives a balance between reliability and informativeness.
- Some consequence relations are provided. The reasoning power of PLP is enhanced with these consequence relations.

# System Structure

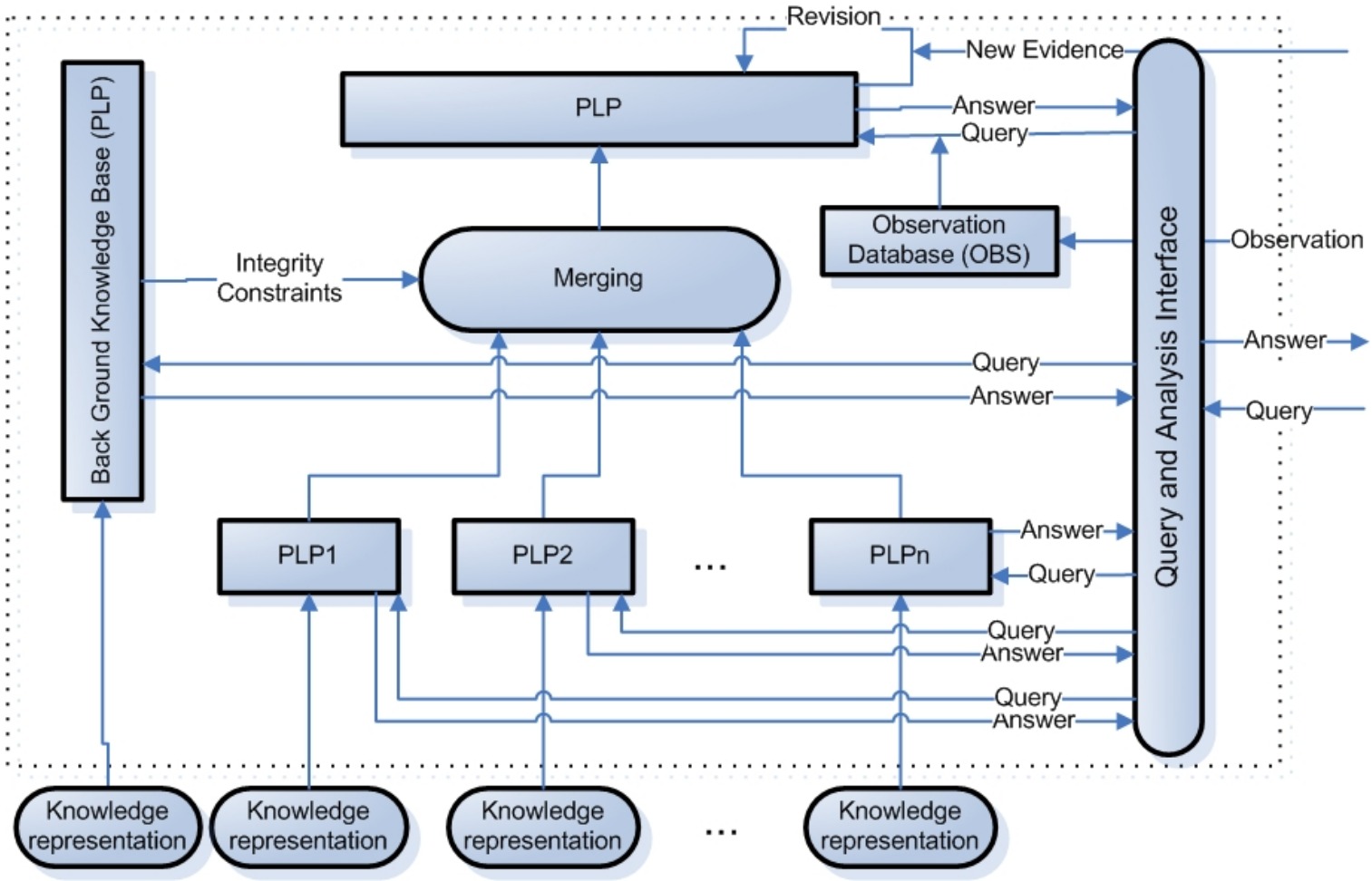


Figure: System Architecture

# Case study I: Rapid sugar kinase enzymes prediction

- From biochemistry on the human enzyme galactokinase.
- Uses galactose as a substrate. Galactose has the molecular formula $C_6H_{12}O_6$.
- $(sub(X)|c1(X,d) \wedge c2(X,u) \wedge c3(X,u) \wedge c4(X,u) \wedge c5(X,u) \wedge c6(X,p))$

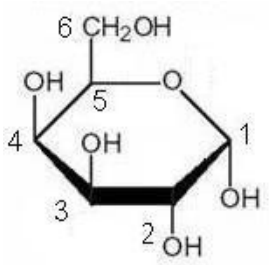


Figure: The $\alpha$-D-Galactose molecule

# Case study I: data

| Sugar Source | C1 -OH | C2 -OH | C3 -OH | C4 -OH | C5 $-CH_2OH$ | C6 -OH | P(substrate) | Product |
|---|---|---|---|---|---|---|---|---|
| Galactose | D | D | U | U | U | P | 1.0 | 1 |
| Glucose | D | D | U | D | U | P | 0.0 | 0 |
| 2-Deoxygalactose | D | A | U | U | U | P | 1.0 | 0.47 |
| Fucose | D | D | U | U | U | A | 0.0 | 0 |
| Talose | D | U | U | U | U | P | [0.4, 0.6] | [0.056,0.084] |
| 4-deoxygluocse | D | D | U | A | U | P | [0, 0.5] | [0,0.021] |
| 3-deoxygalactose | D | D | A | D | U | P | [0.6, 0.9] | [0.036,0.054] |

Table: The compounds and their probabilities and products to be substrates, obtained from published papers.

| Sugar | *C*1 -OH | *C*2 -OH | *C*3 -OH | *C*4 -OH | *C*5 $-CH_2OH$ | *C*6 -OH | P(substrate) | Product |
|---|---|---|---|---|---|---|---|---|
| 2dAll | D | A | D | D | U | P | 0.6529 | 0.4611 |
| 2dGlc | D | A | U | D | U | P | 0.6154 | 0.3939 |
| 2dGul | D | A | D | U | U | P | 0.6694 | 0.5000 |
| I | D | A | A | D | U | P | 0.5869 | 0.4083 |
| II | D | A | A | U | U | P | 0.6676 | 0.5376 |
| 2,3,4d | D | A | A | A | U | P | 0.5509 | 0.4721 |
| 3dAll | D | D | A | D | U | P | 0.6003 | 0.1138 |
| 3dMan | D | U | A | D | U | P | 0.5539 | 0.5000 |
| 3dTal | D | U | A | U | U | P | 0.5636 | 0.4282 |
| III | D | D | A | A | U | P | 0.5321 | 0.3503 |
| IV | D | U | A | A | U | P | 0.5134 | 0.4785 |
| 4dAll | D | D | D | A | U | P | 0.5314 | 0.4611 |
| 4dMan | D | U | U | A | U | P | 0.4706 | 0.4282 |
| V | D | A | D | D | U | A | 0.5463 | 0.4811 |
| VI | D | A | U | D | U | A | 0.5481 | 0.4514 |
| VII | D | A | D | U | U | A | 0.5481 | 0.5000 |
| VIII | D | A | A | D | U | A | 0.5703 | 0.4572 |
| IX | D | A | A | U | U | A | 0.5682 | 0.5020 |
| X | D | A | A | A | U | A | 0.5233 | 0.4814 |
| XI | D | D | A | D | U | A | 0.5451 | 0.3518 |
| XII | D | U | A | D | U | A | 0.5234 | 0.5000 |
| XIII | D | U | A | U | U | A | 0.5278 | 0.4670 |
| XIV | D | D | A | A | U | A | 0.5146 | 0.4179 |
| XV | D | U | A | A | U | A | 0.5064 | 0.4895 |
| XVI | D | D | D | A | U | A | 0.5144 | 0.4811 |
| XVIII | D | U | U | A | U | A | 0.4879 | 0.4670 |

Table: The probabilities and products of some compounds being a substrate by querying on the PLP.

# Case study I: analysis

- The ranking of the compounds in terms of their probability of being a substrate seems mostly reasonable and in line with chemical intuition.
- Overall, the predictions appear to over-estimate the probabilities for each possible substrate.
- The absolute values of the predicted probabilities are less important than the rank order of the compounds.
- The most likely use of such a system is to act as a preliminary screen for potential substrates or inhibitors followed by experimental testing of those compounds.
- Time and expense can be saved if those compounds most likely to be good substrates (or inhibitors) appear at the top of the list and are, therefore, prioritized in the experimental work.

# Case study II: Substrate prediction for NQO1

- NAD(H)-quinone oxidoreductase 1 (NQO1) is a broad specificity enzyme which catalyses the reduction of a range of aromatic compounds.
- It was chosen for the second case study as a large variety of different compounds (including quinones, nitroaromatics and benzimidazoles) have been tested as substrates.
- In contrast to Case study I, the chemical diversity of the known substrates is wider leading to a greater number of variables to consider.

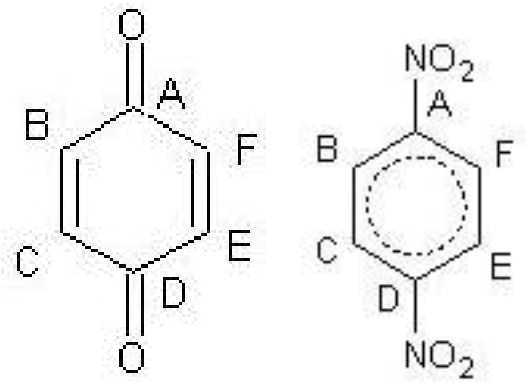


Figure: Examples of NAD(H)-quinone oxidoreductase 1 (NQO1) substrates.

# Case study II: data

| A | B | C | D | E | F | Probability |
|---|---|---|---|---|---|---|
| NO2 | H | H | H | H | H | [0,0] |
| NO2 | H | NO2 | H | H | H | [0,0] |
| NO2 | H | H | CHO | H | H | [0,0] |
| NO2 | NO2 | H | H | H | H | [0,0] |
| NO2 | H | H | NO2 | H | H | [0,0] |
| O | H | H | O | H | H | [0.20,0.28] |
| O | CH3 | H | O | H | H | [0.17,0.31] |
| O | CH3 | H | O | CH3 | H | [0.19,0.33] |
| O | CH3 | CH3 | O | CH3 | H | [0.20,0.28] |

Table: The compounds and their probability intervals, obtained from published papers.

# Case study II: result

| A | B | C | D | E | F | Probability |
|---|---|---|---|---|---|---|
| NO2 | H | H | H | NO2 | H | 0.0000 |
| NO2 | H | H | NO2 | CH3 | H | 0.3194 |
| NO2 | H | H | CHO | CH3 | H | 0.3194 |
| NO2 | H | H | O | CH3 | H | 0.3294 |
| NO2 | H | NO2 | H | CH3 | H | 0.3217 |
| NO2 | H | NO2 | NO2 | H | H | 0.1949 |
| NO2 | H | NO2 | O | H | H | 0.2235 |
| NO2 | NO2 | H | O | H | H | 0.2172 |
| O | H | H | H | NO2 | H | 0.2949 |
| O | H | H | NO2 | CH3 | H | 0.3917 |
| O | H | H | CHO | CH3 | H | 0.3197 |
| O | H | H | O | CH3 | H | 0.3629 |
| O | H | NO2 | H | CH3 | H | 0.4000 |
| O | H | NO2 | NO2 | H | H | 0.3612 |
| O | H | NO2 | O | H | H | 0.3477 |
| O | NO2 | H | O | H | H | 0.3338 |

Table: The Predictions for some compounds.

- the results were broadly similar to those seen in Case Study I

# Conclusion

- Our system can analyze the knowledge contained in PLPs, especially w.r.t. queries.
- Our ignorance provides more information about the underlying knowledge base and is more accurate in terms of reflecting the knowledge in a PLP than some other related works.
- The rank for compounds seems mostly reasonable and in line with chemical intuition.
- Our system can save time and expenses in helping find new substrates.