Integrating and Ranking Uncertain Scientific Data

Landon Detwiler[1], Wolfgang Gatterbauer[2], Brent Louie[1], Dan Suciu[2], Peter Tarczy-Hornoch[1,2]

University of Washington
Seattle, WA, USA

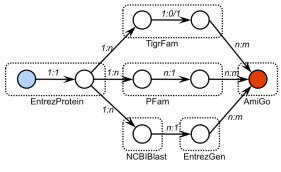
1: Biomedical and Health Informatics
2: Computer Science and Engineering

Motivation

- PROBLEM: Research in biology requires integrating information across heterogeneous data sources. However, biological data contain many uncertainties. As a result, consecutive joins during data integration can lead to many irrelevant results.
- SOLUTION: "BioRank" characterizes uncertainties as probabilistic weights on the integrated data graph and applies probabilistic methods for ranking query results.

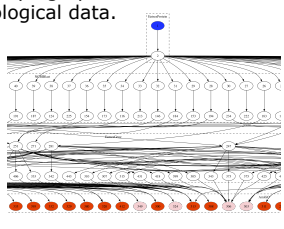
Approach

- We enhanced "BioMediator," an existing mediator-based data integration system that allows exploratory query expansion across different sources.



Example integration schema showing biological data sources, relationships between data entities, and cardinality.

- We construct a probabilistic query graph of the integrated entities and relations. Probabilities in the query graph are derived from "uncertainties" in biological data.



Example expanded query graph. The blue node represents the query, the red nodes are results ranked by relevance (color intensity). Edges represent relations.

- We use domain experts to transform data uncertainties into 4 types of probabilistic weights on both entities and relations on the query graph (p_s, p_r, q_s, and q_r).

Sample p_s and q_s:

	Sets	Records
Entity	$p_s \in [0,1]$	$p_r(a_1, a_2, \ldots) \in [0,1]$
Relationship	$q_s \in [0,1]$	$q_r(b_1, b_2, \ldots) \in [0,1]$

Sample p_r:

EntrezGene

StatusCode	p_r
Reviewed	1.0
Validated	0.8
Provisional	0.7
Predicted	0.4
Model	0.3
Inferred	0.2

Sample q_r: $q_r = -\frac{1}{300}\log(e\text{-}value)$

Examples of probabilistic weights and transformations

- We apply probabilistic scoring methods to calculate a "relevance" score for ranking results. These are based on network reliability and propagation algorithms.

#	*Function (abbr.)*	*GO term*	*r score*
1	sulphonylurea receptor activity	GO:0008281	0.7000
2	potassium ion conductance	GO:0006813	0.7000
3	Interacting selectively with ATP	GO:0005524	0.6999
4	cytoplasmic membrane	GO:0005886	0.6996
5	small-molecule carrier or transporter	GO:0005215	0.6977
...	...	...	...

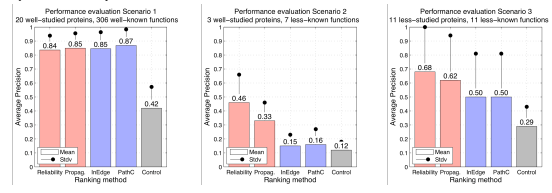
Example of results to a query, ranked by a computed relevance ("r") score.

- We further use several methods that allow efficient evaluation of probabilistic queries.

Experimental Results

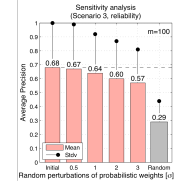
Our motivating application scenario is prediction of protein function. We address 3 questions in our experiments:

- PERFORMANCE: Is probabilistic results ranking better than deterministic approaches such as using neighbor or path-counts to rank nodes? Probabilistic methods clearly perform better for predicting less-known or previously unknown functions.



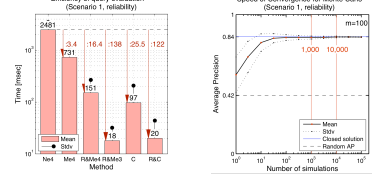
For predicting well-known functions (scenario 1), probabilistic rankings (red) do not perform better than their deterministic counterparts (blue). They clearly do for less-known protein functions (scenarios 2 and 3).

- ROBUSTNESS: BioRank depends on transformation of uncertainties into probabilistic weights. These are used as parameters in BioRank to calculate result rankings (function predictions). How robust are BioRank predictions to systematic variations in these parameter estimates? In a sensitivity analysis they remained very robust.



Random small perturbations to the initial values of BioRank parameters do not negatively affect the quality of rankings across the 3 scenarios.

- EFFICIENCY: How much more expensive are probabilistic rankings? Several methods allowed us to evaluate probabilistic queries in under 20ms in our scenarios.



We could efficiently evaluate queries with the help of 3 methods: (i) an efficient Monte Carlo implementation, (ii) graph reductions, and (iii) a tractable closed solution.

Key Points

- Explicit modeling of uncertainties as probabilities increases our ability to predict less-known or previously unknown protein functions. This suggests that uncertainty models offer utility for knowledge discovery.
- Small perturbations in the input probabilities (parameters) tend to produce only minor changes in the quality of our result rankings. This suggests that probabilistic methods are robust against variations in the way uncertainties are transformed into probabilities.
- Several techniques allow us to evaluate probabilistic rankings efficiently. This suggests that probabilistic query evaluation is not as hard for real-world problems as theory indicates.

Reference

- L. Detwiler, W. Gatterbauer, B. Louie, D. Suciu and P. Tarczy-Hornoch. Integrating and Ranking Uncertain Scientific Data. In Proceedings of the 25th International Conference on Data Engineering, 2009.
- http://www.biomediator.org

Corresponding Author: Wolfgang Gatterbauer, gatter@cs.washington.edu

This work was partially supported by NSF IIS-0513877, NSF IIS-0454425, NSF IIS-0713576, NIH NLM T15 LM07442, and a gift from Microsoft.