Integrating and Ranking Uncertain Scientific Data

Landon Detwiler [1], Wolfgang Gatterbauer [2], Brent Louie [3], Dan Suciu [2], Peter Tarczy-Hornoch [1]

[1]Biomedical and Health Informatics, University of Washington, Seattle, USA
{det,pth}@u.washington.edu

[2]Computer Science and Engineering, University of Washington, Seattle, USA
{gatter,suciu}@cs.washington.edu

[3]Children's Hospital Research Institute, Seattle, USA
brenton.louie@seattlechildrens.org

***Abstract*— Mediator-based data integration systems resolve exploratory queries by joining data elements across sources. In the presence of uncertainties, such multiple expansions can quickly lead to spurious connections and incorrect results. The BioRank project investigates formalisms for modeling uncertainty during scientific data integration and for ranking uncertain query results. Our motivating application is protein function prediction. In this paper we show that: (*i*) explicit modeling of uncertainties as probabilities increases our ability to predict *less-known* or previously *unknown* functions (though it does not improve predicting the well-known). This suggests that probabilistic uncertainty models offer utility for scientific knowledge discovery; (*ii*) small perturbations in the input probabilities tend to produce only minor changes in the quality of our result rankings. This suggests that our methods are *robust* against slight variations in the way uncertainties are transformed into probabilities; and (*iii*) several techniques allow us to evaluate our probabilistic rankings *efficiently*. This suggests that probabilistic query evaluation is not as hard for real-world problems as theory indicates.**

I. Introduction

Much work on managing uncertain data has postulated that uncertainties can be quantified with probabilities and has focused on efficient evaluation techniques for probabilistic databases [1], [2], [3], [4], [5], [6], [7]. However, little has been done to examine the validity of the fundamental assumption that *probabilities are the appropriate model for representing uncertainties in real-world applications*. In this paper, we explore the validity of this assumption and demonstrate how explicitly treating uncertainties as probabilities during data integration improves protein function prediction.

Assigning new functions to proteins is a key challenge in biology. Laboratory experiments, which are still the only reliable means for verifying protein functions, are costly and time consuming. Therefore, in silico techniques that identify and rank the most likely candidates can significantly save resources. At present, the process of identifying likely protein functions requires biologists to perform manual exploratory searches over numerous, heterogeneous online databases. Additionally, there is no clear methodology for combining multiple pieces of evidence and ranking identified candidates. Take as example a researcher who is interested in the functions of the protein ABCC8. She is already aware of its *well-known* functions and is interested in determining if ABCC8 has any additional *less-known* or yet *unknown* functions that relate to diabetes. While biological experiments will ultimately be used to verify predicted functions, she must first discover and suggest them. In her search for candidate functions, she can use many different data sources. Suppose she begins by searching EntrezGene which returns 12 functions. The ABCC8 record in EntrezGene also refers to a record in EntrezProtein. She queries for the related record in EntrezProtein, which contains an amino acid sequence. She can use this sequence as search key in TigrFam. This latter database predicts 53 functions, 3 of which are contained in the 12 results from EntrezGene. Similarly, she can manually follow chains of evidence into additional sources such as NCBIBlast and Pfam. These 4 databases alone lead to 97 candidate functions with varying degrees of evidence. She must now rank those candidates and choose those which are most likely to be correct for further investigation.

In this paper, we present a data integration system that automatically performs exploratory query resolution. We experiment with different ways to model the uncertainty, evaluate the quality of the ranked results, and compare time performance of different query evaluation techniques.

II. A Model for Integrating Uncertain Data

In this section, we describe the data and the query models that allow biologists to explore multiple biological databases in the presence of data and join uncertainties. We describe here only from the mediated schema, since the rest of the integration system (mappings, wrappers, query translations, connection to sources) is based on our previous work [8], [9], [10], [11] (www.biomediator.org).

We use an Entity-Relationship (E/R) model as the mediated schema between data sources. An entity set has a schema $P(\underline{id}, a_1, a_2, \ldots)$ where id is the key, and a relationship has a schema $Q(\underline{id, id'}, b_1, b_2, \ldots)$ where id, id' are foreign keys to two entity sets P, P' that Q relates. Here, $a_1, a_2, \ldots, b_1, b_2, \ldots$ denote attributes. Every data source that we integrate exports one ore more entity sets. Our system computes a number of relationships between the sources to achieve the actual integration, e.g. by following foreign keys, looking up aliases, or even matching keywords.

Our method requires transforming uncertainties of various kinds into *probabilistic weights* between 0 and 1. These weights are not probabilities in a statistical sense, but are better interpreted as subjective and *relative weights* of evidence

which help us compute relevance scores and rank query results. In this process, they are formally treated as probabilities. We will empirically demonstrate the utility of this approach.

There are 4 types of probabilistic weights in our system, denoted generically p_s, q_s, p_r, and q_r, with the following meaning [12]: p_s represents our degree of confidence in an entire entity set P, while q_s represents our degree of confidence in an entire relationship Q; p_r represents our confidence in a particular record in P and is computed by a function of the record's attributes, $p_r(a_1, a_2, ...)$; finally, q_r represents the degree of confidence in a concrete relationship in Q, and is also computed by a function. Some functions are modeled as lookup tables where certain attributes like "evidence codes" are mapped to weights between 0 and 1. Other uncertainty attributes like "e-values", that are continuous and non-linear, are transformed on a logarithmic scale into probability weights. Thus, our data model is a probabilistic database, where each schema component and each data record has a probabilistic weight representing the confidence in that item. The actual weights and transformation functions were determined in extensive discussions with our collaborators at Seattle Children's Hospital Research Institute (SCHRI). This raises the question of how sensitive the system's performance is with regard to variations in the assigned weights. In Section IV we will show that it is very robust.

Conceptually, we represent the entire integrated data as a *probabilistic data graph*, which is a labeled, directed graph $G = (N, E, p, q)$, where N is the set of nodes, $E \subseteq N \times N$ the set of edges, and $p : N \rightarrow [0, 1]$ and $q : E \rightarrow [0, 1]$ are probability labels for each node and edge, respectively. In the mapping from the E/R schema, data records become nodes and relationships edges. The node and edge probabilities $p(i) = p_s(i) \cdot p_r(i)$ and $q(i, j) = q_s(i, j) \cdot q_r(i, j)$ are derived by multiplying the respective set with record probabilities.

BioRank supports a simple, yet powerful class of queries, which we call *exploratory queries*. A user creates a query or source node s by selecting an input entity set P_{i}, one of its attributes *attr* and a value, and a set of output entity sets $\{P_{\text{o}1}, \ldots, P_{\text{o}n}\}$: $(P_{\text{i}}.attr = \text{"value"}, \{P_{\text{o}1}, \ldots, P_{\text{o}n}\})$. The system retrieves all records in P_{i} whose attribute *attr* matches the value, then follows all links recursively to find all reachable records. It, thus, constructs a *probabilistic query graph* $G = (N, E, p, q, s, A)$ where $s \in N$ is the source node, and $A \subset N$ the answer set with $A = \{v \mid \exists x \in P_{\text{i}}, \text{"value"} \in x.attr, x \rightarrow v, \exists i : v \in P_{\text{o}i}\}$, were $x \rightarrow v$ means that there exists a path from node x to node v in the entity graph. A *relevance score* $r : A \rightarrow R$ imposes a partial order on the answer set by assigning each node in the answer set $t \in A$ a relevance score $r \in R$, where R stands for the range. The result is a ranked answer set of records which can be reached from the query node.

We illustrate the working of our system within the context of our motivating application and with Fig. 1. In response to the query (EntrezProtein.*name* = "ABCC8", {AmiGO}), the system creates a new query node s, then links to all records in the input entity set EntrezProtein whose attribute

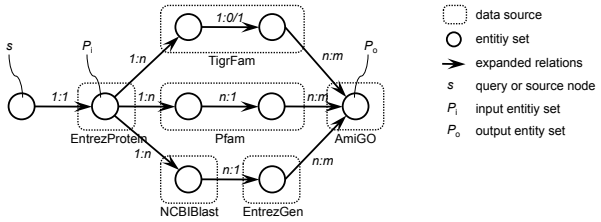

Fig. 1. Subset of the E/R schema used in our motivating application.

name has value "ABCC8". From here the system continues in three separate sources, searching for all possible paths to some record in the output entity set AmiGO. Note that the figure illustrates the graph only at the schema level: the probabilistic data graph has 684 nodes, 977 edges, and the answer set consists of 97 individual nodes. Many of these 97 answers are wrong, which is a typical problem for exploratory queries; uncertainty and imprecision in the data lead to many incorrect answers. To address this issue, BioRank evaluates the level of certainty in each result, assigns each a relevance score and presents the user with a ranked list of protein functions.

III. Ranking Integrated Data

Previous work [13] proposed to use simple graph metrics like number of incoming edges and path length from a query to a target node to rank connected information in a biological data integration system. In contrast, we explored several relevance functions that explicitly take the uncertainties at each integration step into account. The two most important ones are *reliability*, which follows a possible world semantics, and *propagation*, which is a variation using an independence assumption. Deterministic ranking methods are *inEdge*, which counts the number of incoming edges to a target node, and *pathCount*, which counts the number of different paths between the query and a target node. We discuss these methods in more detail in [14]. Here we focus on reliability.

The *reliability semantics* calculates relevance scores by interpreting the query graph as a network reliability problem [15]. More specifically, we use the generalized source-target reliability problem with node failures that can be reduced to the standard network reliability problem by removing node failures and reifying the graph. Input is a *probabilistic query graph* $G = (N, E, p, q, s, A)$ where $p : N \rightarrow [0, 1]$ and $q : E \rightarrow [0, 1]$ are the probabilities $p(n)$ and $q(e)$ that a node or edge is present, $s \in N$ the source node, and $A \subset N$ the answer set. For each target node $t \in A$, the reliability score $r(t)$ is then the probability that t is connected to s and active. We use these relevance scores to rank the nodes in the answer set. Our rationale for the reliability semantics is that it is equivalent to the *possible worlds semantics* in probabilistic databases [5]. We developed three techniques to make query evaluation for the reliability semantics tractable in our setting: (*i*) efficient *Monte Carlo simulations*, (*ii*) repeated *graph reductions*, and (*iii*) *tractable closed solutions*. We cover them in detail in [14]. In the next section, we evaluate their performance improvements.

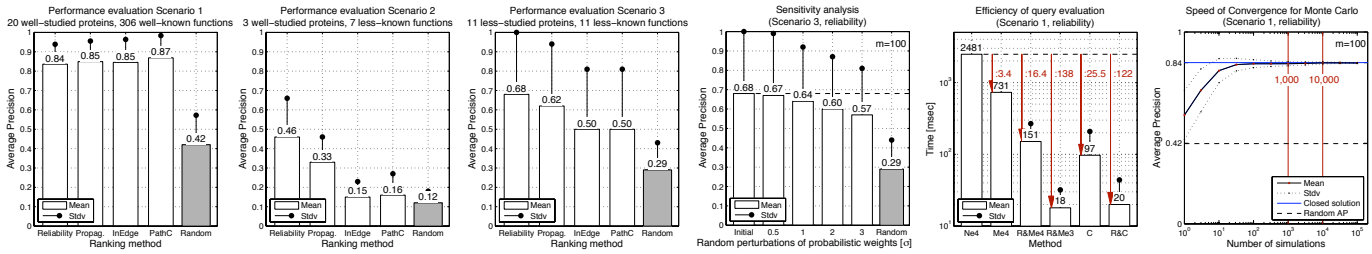

Fig. 2. (a): For well-known functions, inEdge and pathCount give slightly better rankings than probabilistic scores reliability and propagation. (b,c): The later, however, perform clearly better for less-known functions and on less-studied genes. (d): Probabilistic rankings remain robust over a wide range of random perturbations of the probabilistic weights. (e): Combination of several techniques reduce required time for probabilistic query evaluation by 2 orders of magnitude (Ne4: Naive Monte Carlo 1e4 sim., Me4: Efficient Monte Carlo 1e4 sim., R&M: Graph reductions prior to Monte Carlo, C: Tractable closed solution alone, R&C: Graph reductions prior to closed solution). (f): Reliability rankings already converge around 1e3 trials in the Monte Carlos simulations.

IV. Experiments

The output of BioRank is an ordered list of predicted functions for a query protein. To consistently assess ranking performance of all four scoring methods across different scenarios, we use the measure *average precision* (AP) and the analytic method proposed in [16] that accounts for ties. As benchmark, we use the expected average precision of an arbitrarily ranked result list. We call this measure *random average precisions* (AP_{rand}) and define it as the expected AP when a list of n total items with k relevant is randomly sorted.

Scenario 1: Well-known functions for well-studied proteins. Here we chose 20 well-studied proteins with their 306 *well-known* functions from the iProClass database. These functions are a good reference standard as they have highly reliable evidence (over 46% of functions in our test set are confirmed by experiments versus 5% for proteins in general). The data sources we integrated were Pfam, TIGRFAM, NCBIBlast and Entrez. iProClass was not incorporated as it was kept as reference set. Figure 2a shows that all 4 ranking methods perform significantly better than random sorting of predicted function; the two deterministic ones, inEdge and pathCount, perform slightly better than reliability and propagation.

Scenario 2: Less-known functions for well-studied proteins. Given the 20 reference proteins, we searched for *less-known* functions that were not yet described in their iProClass record. Before being entered into databases and becoming de facto well-known, newly discovered protein functions are first described in publications. Thus, we manually searched through PubMed for recent publications on the reference proteins and could validate 7 predictions from BioRank for 3 of the 20 proteins. Figure 2b lists AP just for the 7 newly found functions. This time, the probabilistic ranking functions performed visibly better than the deterministic ones.

Scenario 3: Unknown functions for less-studied proteins. This scenario models the problem of assigning function to hypothetical proteins (proteins of unknown function). Using a manual approach described in [17], biological experts created a reference set of 11 proteins with one function each. The reference set is very small due to the procedure's high cost in terms of human effort. Figure 2c shows that reliability and propagation clearly perform better than deterministic rankings.

Sensitivity analysis. Section II describes the transformation of uncertainties into probabilistic weights. An important question is how can we guarantee that the uncertainty transformations proposed by domain experts are actually correct. Put differently, how sensitive are the predictions of BioRank to variations in the way probabilistic weights are estimated? To answer this question, we performed a multi-way sensitivity analysis of the ranking quality of our probabilistic methods with respect to systematic perturbations of all probabilistic node and edge weights simultaneously. Specifically, we used a method proposed by [18] to add random noise at $\sigma = 0.5, 1, 2, 3$ standard deviations to initial values, averaged over $m = 100$ repeated experiments, and compared AP of our probabilistic rankings across the previously described 3 scenarios. Figure 2d shows that the ranking quality (here for reliability) does not significantly decrease until noise of $\sigma = 3$ is added. Also, for less known information, AP remains still higher than for the deterministic alternatives (compare with Fig. 2c). This interesting result suggests that our probabilistic rankings are very robust against subjective and slightly varying quantifications of probabilistic weights and, hence, our decision to *let domain experts perform those transformations* is well-justified. It is also consistent with observations in AI that probabilistic belief networks often show a similar robustness to imprecise input probabilities [19].

Efficiency of query evaluation. Section III mentions and [14] describes in detail several techniques for speeding up probabilistic query evaluation. Here, we evaluate our techniques for the reliability semantics on the 20 query graphs for scenario 1 (Fig. 2e). • By limiting the amount of necessary node and edge simulations, our *efficient Monte Carlo* (MC) implementation achieves an average speed-up of 3.4 over the naive implementation. • Our 20 original graphs have on average 520 nodes and 695 edges. Applying our *graph reductions*, we can reduce their number by an average factor of 4.5. Combining those reductions with subsequent efficient MC, we improve the speed-up to a factor of 16.4. • We developed an indicator when reliability rankings can be calculated efficiently in a *tractable closed form*. Given our setup from Fig. 1, the theory predicts

that the queries can be reduced by just evaluating the subtrees to each answer node sequentially by themselves. Our theory proves to be correct and useful. The closed solution alone achieves a speed-up of 25.5; combining graph reductions and closed solutions improves that to a factor of 122. • To study the *convergence of MC*, we run it for a varying number of simulation steps and calculated mean and standard deviation averaging over $m = 100$ runs each. Figure 2f shows that already 1,000 trials achieve high average ranking accuracy. Combining 1,000 trials with efficient MC, the speed-up is around 138 over a naive MC with 10,000 trials.

Overall, several techniques allow our system to improve probabilistic query evaluation by over 2 orders of magnitude over a naive implementation; query times are also within 1-2 orders of magnitude of deterministic rankings, which is a very positive result.

V. Related Work

There has been recent interest in developing general-purpose query evaluation methods on *probabilistic databases* [1], [2], [3], [4], [5], [7]. Probabilities are typically associated at the tuple level, and the query language is a subset of SQL. Probabilistic data is used in data integration in [20], [21], [6]. These studies start from the assumption that the data is probabilistic, or assume that the data is uncertain but can be modeled as probabilistic data.

Approaches have been proposed that exploit the global link structure between integrated data items for *ranking biological entities*, for example using path length and inEdge cardinality [13]. Biozon [22] uses a algorithm similar to PageRank. However, uncertainty of individual data points are not taken into account for determining ranks. BioRank is unique in that it not only accounts for the link structure between data items, but also explicitly models uncertainty inherent in data entities, links between entities, and at the data source level.

Network reliability and propagation algorithms have been proposed for inferring protein complex membership [23], [24]. These studies have shown promise in terms of inferring biological information from a network of data. They differ from our approach in that they create and evaluate a specific model and are not concerned with general-purpose data integration and information retrieval uncertainty semantics.

VI. Conclusions

The high-level take-away from our work is that while probabilistic approaches are not necessary and valuable for all data integration problems, they are ideally suited for highly uncertain domains such as new scientific knowledge discovery. In particular, we showed: (*i*) Enriching an existing biological data integration application allowed us to consistently improve ranking performance on *less-known* information, not however for already *well-known* information; (*ii*) Our probabilistic rankings were very robust against subjective and slightly varying estimates of domain experts; and (*iii*) we gave several efficient methods to perform probabilistic ranking.

VII. Acknowledgment

We thank Eugene Kolker and Peter Myler for their ongoing collaboration, and Nilesh Dalvi for his useful comments. This work was partially supported by NSF IIS-0513877, NSF IIS-0454425, NSF IIS-0713576, NIH NLM T15 LM07442, and a gift from Microsoft.

References

[1] N. Dalvi and D. Suciu, "Efficient query evaluation on probabilistic databases," in *Proc. VLDB*, 2004.
[2] J. Widom, "Trio: A system for integrated management of data, accuracy, and lineage," in *Proc. CIDR*, 2005.
[3] C. Re, N. Dalvi, and D. Suciu, "Efficient Top-k query evaluation on probabilistic data," in *Proc. ICDE*, 2007.
[4] P. Sen and A. Deshpande, "Representing and querying correlated tuples in probabilistic databases," in *Proc. ICDE*, 2007.
[5] N. Dalvi and D. Suciu, "Management of probabilistic data: foundations and challenges," in *Proc. PODS*, 2007.
[6] X. Dong, A. Halevy, and C. Yu, "Data integration with uncertainty," in *Proc. VLDB*, 2007.
[7] L. Antova, T. Jansen, C. Koch, and D. Olteanu, "Fast and simple relational processing of uncertain data," in *Proc. ICDE*, 2008.
[8] P. Mork, A. Y. Halevy, and P. Tarczy-Hornoch, "A model for data integration systems of biomedical data applied to online genetic databases," in *AMIA Annual Fall Symposium*, 2001.
[9] R. Shaker, P. Mork, M. Barclay, and P. Tarczy-Hornoch, "A rule driven bi-directional translation system remapping queries and result sets between a mediated schema and heterogeneous data sources," *Jour Amer Med Inform Assoc, Fall Symposium Suppl*, 2002.
[10] L. Donelson, P. Tarczy-Hornoch, P. Mork, C. Dolan, J. Mitchell, M. Barrier, and H. Mei, "The BioMediator system as a data integration tool to answer diverse biologic queries," in *Medinfo*, 2003, pp. 768–72.
[11] R. Shaker, P. Mork, J. Brockenbrough, L. Donelson, and P. Tarczy-Hornoch, "The BioMediator system as a tool for integrating databases on the web," in *Proc. IIWeb*, 2004.
[12] B. Louie, T. Detwiler, N. Dalvi, R. Shaker, P. Tarczy-Hornoch, and D. Suciu, "Incorporating uncertainty metrics into a general-purpose data integration system," in *Proc. SSDBM*, 2007.
[13] Z. Lacroix, L. Raschid, and M.-E. Vidal, "Efficient techniques to explore and rank paths in life science data sources," in *Proc. DILS*, 2004.
[14] L. Detwiler, W. Gatterbauer, B. Louie, D. Suciu, and P. Tarczy-Hornoch, "Integrating and ranking uncertain scientific data," University of Washington, Tech. Rep., 2008.
[15] C. Colbourn, *The Combinatorics of Network Reliability*. New York, NY, USA: Oxford University Press, 1987.
[16] F. McSherry and M. Najork, "Computing information retrieval performance measures efficiently in the presence of tied scores," in *Proc. ECIR*, 2008.
[17] B. Louie, P. Tarczy-Hornoch, R. Higdon, and E. Kolker, "Validating annotations for uncharacterized proteins in Shewanella oneidensis," 2008, (manuscript in preparation).
[18] M. Henrion, M. Pradhan, B. Favero, K. Huang, G. Provan, and P. O'Rourke, "Why is diagnosis using belief networks insensitive to imprecision in probabilities?" in *Proc. UAI*, 1996.
[19] M. Pradhan, M. Henrion, G. Provan, B. Del Favero, and K. Huang, "The sensitivity of belief networks to imprecise probabilities: an experimental investigation," *Artificial Intelligence*, vol. 85, no. 1-2, pp. 363–397, 1996.
[20] A. Nierman and H. Jagadish, "ProTDB: Probabilistic data in XML," in *Proc. VLDB*, 2002.
[21] M. van Keulen, A. de Keijzer, and W. Alink, "A probabilistic XML approach to data integration," in *Proc. ICDE*, 2005.
[22] A. Birkland and G. Yona, "BIOZON: a system for unification, management and analysis of heterogeneous biological data," *BMC Bioinformatics*, vol. 7, p. 70, 2006.
[23] S. Asthana, O. D. King, F. D. Gibbons, and F. P. Roth, "Predicting protein complex membership using probabilistic network reliability," *Genome Res.*, vol. 14, no. 6, pp. 1170–1175, 2004.
[24] J. Weston, A. Elisseeff, D. Zhou, C. S. Leslie, and W. S. Noble, "Protein ranking: from local to global structure in the protein similarity network," *PNAS*, vol. 101, no. 17, pp. 6559–6563, Apr. 27 2004.