Link Discovery in Graphs Derived from Biological Databases
(Research Paper)

Petteri Sevon, Lauri Eronen, Petteri Hintsanen,
Kimmo Kulovesi, and Hannu Toivonen*

HIIT Basic Research Unit, Department of Computer Science,
P.O. Box 68, FI-00014 University of Helsinki, Finland
{Petteri.Sevon, Lauri.Eronen, Petteri.Hintsanen, Kimmo.Kulovesi,
Hannu.Toivonen}@cs.helsinki.fi

Abstract. Public biological databases contain vast amounts of rich data that can also be used to create and evaluate new biological hypothesis. We propose a method for link discovery in biological databases, i.e., for prediction and evaluation of implicit or previously unknown connections between biological entities and concepts. In our framework, information extracted from available databases is represented as a graph, where vertices correspond to entities and concepts, and edges represent known, annotated relationships between vertices. A link, an (implicit and possibly unknown) relation between two entities is manifested as a path or a subgraph connecting the corresponding vertices. We propose measures for link goodness that are based on three factors: edge reliability, relevance, and rarity. We handle these factors with a proper probabilistic interpretation. We give practical methods for finding and evaluating links in large graphs and report experimental results with Alzheimer genes and protein interactions.

1 Introduction

The amount of publically available biological data is growing at a tremendous pace, as new information about genomes, proteomes, interactomes etc. is published daily. Despite the large amount of that information, it is clear that it only represents a tiny fraction of the biological knowledge that potentially will be discovered. For instance, consider the functions of genes: in the Gene Ontology database[1], 29.5% of those gene products that have an annotation for a molecular function, the annotation at the time of writing is "unknown". This example only represents some of the facts we know that we do not know yet.

We present novel computational methods for predicting some of the missing information, with the primary aim of producing and ranking new biological hypothesis for life scientists working on their own specific problems. We assume

* Work done while visiting the University of Freiburg.
[1] http://www.godatabase.org

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 35–49, 2006.

a fairly simple and generic form for the input data: a graph where biological entities and concepts constitute the set of vertices, and the edges correspond to known and annotated relationships between the vertices. In this framework, a yet undiscovered link between two entities or concepts may be manifested as a path or a subgraph connecting the corresponding vertices. Qualitative hypotheses for the biological mechanisms are generated by discovering such paths or subgraphs. In this paper, we use the term *link* to refer to any connections between two vertices in the graph, potentially output as a hypothesis for a biological relation.

Not all paths represent a biologically meaningful links. Two edges incident on a vertex may constitute a spurious path, or edges may not be completely reliable. To be able to address more interesting questions, such as evaluation of the statistical significance of a link, or ranking a set of vertices in order of strength of linkage to a given vertex, we need a way of quantifying the strength of a link. This will be a central topic of this paper.

In our scenario for the analysis, a life scientist poses queries to a graph database system. In a simple form, such a query can ask if a path exists between two given concepts, and how strong the link is. In a more complex setting, the user may submit sets of vertices and ask the system to find, evaluate and rank subgraphs connecting any pair of given vertices.

As a motivating example, consider gene mapping for a particular phenotype. The mapping may have resulted in a large set of candidate genes. When further expensive analyses are planned for the wet lab, the investigators first compare the candidates in the light of what is known about them in the public databases and literature, hoping to be able to concentrate the efforts and resources on the most promising candidates. Due to the lack of automated methods, the work is mostly done by manually browsing the databases. This is a slow and laborious process, and necessarily limits the extent and coverage of the search. Our methods aim at partial automation of such tasks. As for the specific example, methods for automated discovery and analysis of connections between a candidate gene and a phenotype have only recently started to emerge [1,2].

In this paper, we propose a method for measuring the strength of a link based on the two-terminal network reliability [3] between the end vertices. The main contributions of the paper are a novel application of the network reliability measure, as well as a unique way of assigning probabilities to the edges based on three aspects: reliability, relevance, and rarity. Reliability reflects the confidence to the data source, relevance is a subjective measure of importance, and rarity rewards (informative) edges between nodes with low degrees. We give methods for finding good paths and subgraphs and for evaluating their quality. The applicability of the methods is not restricted to gene–phenotype links; they can be used for analyzing the link between any pair of concepts, and potentially even in completely different application areas.

Related work. Our work can be characterised as link discovery (link mining, see, e.g., [4] for a review)—or, more specifically, as link prediction; we aim at predicting links between pairs of vertices, where none exist in the form of direct edges. We work on the abstract level of graphs. This gives our methods the