Auto-Extraction, Representation and Integration of a Diabetes Ontology using Bayesian Networks

Ken McGarry[*][†], Sheila Garfield[*] and Stefan Wermter[*]
[*]School of Computing and Technology, University of Sunderland, UK
[†]School of Pharmacy, University of Sunderland, UK
{ken.mcgarry,sheila.garfield,stefan.wermter}@sunderland.ac.uk

Abstract

This paper describes how high level biological knowledge obtained from ontologies such as the Gene Ontology (GO) can be integrated with low level information extracted from a Bayesian network trained on protein interaction data. We can automatically generate a biological ontology by text mining the type II diabetes research literature. The ontology is populated with the entities and relationships from protein-to-protein interactions. New, previously unrelated information is extracted from the growing body of research literature and incorporated with knowledge already known on this subject from the gene ontology and databases such as BIND and BioGRID. We integrate the ontology within the probabilistic framework of Bayesian networks which enables reasoning and prediction of protein function.

1 Introduction

The large amounts of genomic and proteomic data that are generated by biological experiments is now enabling deeper insights into cellular and molecular function. New technologies such as microarrays and electrophoresis gels are providing vast quantities of experimental data at unprecedented rates. All of this information needs to be stored and carefully annotated. With each new experiment providing details of new protein-to-protein interactions, new biological pathways and new genes it is essential that these discoveries are made available to the scientific community. To this end, online scientific databases are now in place that disseminate these results. These databases such as the popular Gene Ontology (GO) are updated at intervals to reflect the latest developments [1]. The updating is done by experts who manually revise each entry by reading the research literature and annotating the database collections accordingly. Unfortunately, hand annotation is a slow process and the databases are lagging behind the experimental work by a considerable margin.

Our particular research area is that of diabetes, in particular the effects of insulin resistance on protein expression and insulin regulated protein trafficking in fat cells. In recent years there has been a dramatic worldwide increase of those suffering with diabetes. In the year 2000, there were 171 million cases and by 2030 the World Health Organization (WHO) has predicted there will be 366 million people suffering from this condition ($www.who.int/diabetes/facts/$). The WHO data is for diagnosed cases but the undiagnosed cases are estimated by the WHO at 14.6 million alone for the US.

In this paper we present our results of how we automatically generate a viable ontology based on information extraction of keywords from the research literature. The keywords define the entities and relationships of important genes, gene relationships, protein-to-protein interactions operate and co-exist in biological processes related to insulin resistance. Furthermore, the ontology is cast within a probabilistic framework using Bayesian networks which are used for the inferencing and prediction of protein function. The remainder of this paper is structured as follows; section two outlines our information extraction scheme for identifying the entities and relationships of interest, section three provides an overview of biological ontologies and gives details of how we use Bayesian networks for inference and reasoning. Section four discusses our methodology and experimental results, section five reviews the related work and our claim for novelty and finally section six presents the conclusions.

2 Biological Ontologies and Bayesian Networks

In this section we briefly motivate the need for ontologies and define their limitations with respect to the biological field and for knowledge discovery. Ontologies describe the concepts and relationships that exist for a particular area of interest. They are very useful for the semantic labeling of concepts or definitions [5, 2]. This process ensures that entities which are equivalent to other entities in separate databases are identified as referring to the same concepts. Even if these entities have different names or forms they can still be identified by semantic labeling. The role of semantics therefore is much deeper than matching the co-occurrence of a tag or label, since it defines the relationship that exists between concepts.

The use of ontologies in biology for the semantic integration of heterogeneous data is receiving increased attention, however problems occur because of the dynamic, changing nature of biological knowledge [7]. These difficulties arise from the highly complex structures that are expensive and problematic to update and maintain [3]. Another, related problem is that current ontologies have a rather limited vocabulary and cannot express the richness of biological information. Little attention has been paid to defining the *relations*, much of the research effort and complexity of structure has concentrated on defining the *terms*. Other considerations that are important are the spatial and temporal characteristics of the entities.

Furthermore, ontologies such DAML+OIL, OWL and RDF are based on crisp logic and have difficulty managing uncertainty; incomplete data and noisy information that is encountered in many domains, especially the bioinformatic field. Our research is concerned with Type 2 diabetes, in order to develop a suitable ontology it is necessary to identify the relevant entities within the domain, their attributes and the relationships that exist between these entities.

2.1 Bayesian networks for Ontology Inference and Integration

Ontologies are experiencing increased interest as they are perceived by many as a mechanism for the unification of biology [7]. Within this framework the Bayesian approach can be seen as both a learning mechanism and as a knowledge representation technique [8].

Bayes theorem is shown in equation 1 and presents the probability of the hypothesis (H) conditionalised on evidence (E).

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E \mid H)P(H) + P(E \mid \neg H)P(\neg H)} \tag{1}$$

where: $P(H \mid E)$ defines the probability of a hypothesis conditioned on certain evidence, $P(E \mid H)$ is the likelihood, $P(H)$ is the probability of the hypothesis prior to obtaining any evidence, is the $P(E)$ evidence. Therefore, according to Bayesian theory we can update our beliefs regarding the hypothesis when provided with new evidence that is conditional upon using probabilities and is called *conditionalization*.

The conditional probability distributions (CPD) are described by $P(X_i \mid U_i)$, where X_i represents node i and U_i are its parent nodes. We must specify the prior probabilities of the nodes and the conditional probabilities of the nodes given all the combinations of their ancestor nodes. The joint distribution of random variables is given by $X = \{X_1, ..., X_n\}$ and together with the CPD values is used to calculate the choice of X_i and is given by :

$$P(X_1, ..., X_n) = \prod_i P(X_i \mid U_i) \tag{2}$$

The CPD's values are easy enough to calculate and inference but require the number of parameters is dependent upon the number of parent nodes, they are usually represented in table format. The nodes are assumed to be discrete or categorical values, however, continuous values may be discretised [6].

$$P(X_1, ..., X_n) = \frac{1}{Z} \prod_j \pi_j[C_j] \tag{3}$$

3 Methods and Results

We reviewed the literature associated with Type 2 diabetes, the initial focus associated with protein interaction in diabetes and from this review a list of "events" indicative of protein interactions was identified, eg, activate, inhibit and modulate. This list was used as the starting point to help identify which entities are involved in each type of action or relation. After identifying the names of possible event relations the focus moved to identifying potential entities involved in these relations. In order to complete this task a suitable dataset was required. A search of the PubMed database was conducted and 6113 abstracts, related to Type 2 diabetes were used; this dataset is used throughout each subsequent stage of this work. Initially a count was made of the number of times each of the action words occurred in this sample dataset. Some of the words, eg, "acetylate" and "destabilize" did not occur at all, while other words such as "interaction" and "suppression" occurred more frequently.

We now explain how the various parts of our system function together, the information extraction technique synthesizes the entities and relationships from the literature abstracts and generates the structure for a specific ontology on insulin resistance. We then use the ontologies structure to build a Bayesian network for the purposes of inference and prediction of new protein-to-protein interactions. The relative frequencies of the keywords (entities and relationships) are used to construct the conditional probability tables which define the parent/child node relationships.

3.1 The Extracted Ontology and Bayesian network Mapping

Initially, one of these action words, "interaction" was selected to identify possible entities involved in a relation. The word "interaction" however generally forms part of a phrase such as "interaction between", "interaction of", and "interaction with", and therefore each of these phrases would be used by the algorithm to search for potential entities. The first phrase used was "interaction between". Examples of the resulting phrases extracted are provided in the table 2.

Ultimately, the successful application of Bayesian techniques is dependent on the use of *prior knowledge* to improve the estimation of the posterior. If a prior belief exists about a situation then

Table 1. Biological keywords

Action Word	No	Action Word	No	Action Word	No
acetylate	0	inhibit	109	phosphorylates	5
acetylated	1	inhibited	95	phosphorylation	362
acetylates	0	inhibition	222	regulate	62
acetylation	0	inhibits	59	regulated	62
activate	47	interact	34	regulates	35
activated	69	interacted	0	regulation	333
activates	18	interacting	14	stabilization	6
activation	435	interaction	213	stabilize	3
bind	31	interactions	101	stabilized	3
binding	914	interacts	7	stabilizes	3
binds	16	modulate	74	suppress	56
destabilizes	0	phosphorylated	15		

Table 2. Biological keywords extracted for the ontology for the phrase “interaction between”

Preceding word	Following words
the	thyroid function and insulin sensitivity
the	dysregulated fat and glucose metabolism
strong	insulin resistance and serum
significant	obesity and insulin resistance
possible	BMI and the adiponectin gene

we can use this information to pre-structure our BN. For example if a particular gene (IPA) is known to regulate several target genes (GDH, GL4, HK2), we would then assign this relationship within the BN by setting the edges between these two entities and setting the values in the conditional probability table to define the structural prior accordingly. This is a powerful strategy, but only when it makes sense to do so. The application of incorrect beliefs will produce unreliable estimates of the true posterior regardless of the abundance of the likelihood evidence. Equation 4 shows how we modify the BN with prior knowledge (causal intervention) from the extracted ontology [4].

$$P(X_{i,j} = z \mid par_M(x), M, \theta : X_{i,j} = Z, ...) = 1 \tag{4}$$

where par_M are the parameters within the model, $X_{i,j}$ are the known effects of the parents of a given node, θ is the conditional probability conditionalized and represents the causal conditions. The biological knowledge is incorporated into the BN by specifying the probability for the existence of each potential connection (edge) between them. We assume independence between edges and the variables in the BN are also assumed to be discrete, this ensures that the calculations are computationally tractable.

Figure 1 shows the structure of a section of our ontology. The nodes are the entities and the arcs determine the relationships between them. The numbers in brackets preceded by “GO:” are the probabilities of the term occurring in the GO ontology, the numbers.

For example the following abstract fragment captures knowledge about several proteins and their interactions:

“Overexpression of the cytosolic domain of syntaxin 6 **did not** *affect insulin-stimulated glu-*

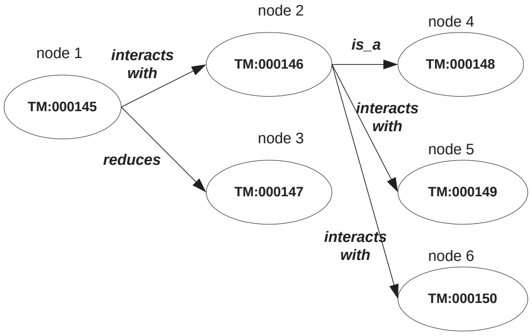

Figure 1. Fragment of the ontology (entities and relations) extracted from the literature

cose transport, but increased basal deGlc transport and cell surface Glut4 levels. Moreover, the syntaxin 6 cytosolic domain significantly reduced the rate of Glut4 reinternalization after insulin withdrawal and perturbed subendosomal Glut4 sorting; the corresponding domains of syntaxins 8 and 12 were **without** *effect."*

We encountered difficulties with negative implications, i.e. the "did not" and "without effect" phrases negate the occurrence of the relationship but would be taken by the information extraction algorithm as a positive relationship. A more elaborate NLP technique or further crafting of specific regular expression templates would reduce this effect.

3.2 Validation against Existing Knowledge

We determined a base line accuracy for our system by "rediscovering" known protein-to-protein interactions from the literature and validating the relationships through accessing a number of online database and ontology repositories. The most up to date and complete is the gene ontology (GO), we compare extracted relationships from our ontology with the GO structure. To determine the accuracy, we apply the well known information retrieval measures of recall and precision. We define recall as the percentage of entity relations represented in the GO and correctly identified. We define precision as the the percentage of relations found in GO and returned by our system.

The recall and precision are calculated by:
$recall = TP/(TP + TN)$, $precision = TP/(TP + FP)$, where: TP=true positives such as , FP= false positives, TN= true negatives and FN= false negatives.

Table 3. Recall and Precision of IE on protein-to-protein interaction data

Keyword	TP	TN	FP	FN	Recall	Precision
interact	100	171	20	32	37	83
bind	200	167	17	14	54	92
promote	240	188	17	15	56	93
inhibit	230	178	12	19	56	95

We should note that certain errors in GO have been identified, inconsistencies and even spelling mistakes. We have also identified that certain GO terms are too general and a more specific term would have been more appropriate. Thus entries with low semantic similarity but high functional similarity can be identified. The GO ontology structure is extremely limited with total reliance on $''is\_a''$ type links. This means that a large amount of semantic information that was originally available from the research articles is missing. We suspect that as ontologies such as GO increase in the number of entities, the relationships between will take on increased value. However, without incorporating the semantic similarity of the entities any increase in size will reduce the ontology to free text.

4 Conclusions

We have demonstrated that the integration of low level genomic data is possible with the higher order structures found in text by mapping them through an ontology. This process is critically dependent on the level of granularity used. We use Bayesian networks to learn from data but also to map existing ontological relations to new Bayesian network structures. Clearly, further work is needed, however, we have extended the current knowledge of automatically generating and integrating ontologies from low level data. The utilization of ontologies as a framework for guiding the knowledge discovery process has to date received little attention. The experimental results presented in this paper led us to conclude that a principled approach such as the Bayesian framework can successfully integrate and represent heterogeneous data and knowledge.

5 Acknowledgements

This work was part supported by a Research Development Fellowship funded by HEFCE and the Biosystems Informatics Institute (Bii).

References

[1] M. Ashburner. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[2] J. Bard and S. Rhee. Ontologies in biology: design applications and future challenges. *Nature Reviews Genetics*, 5:213–222, 2004.

[3] C. Blaschke and A. Valencia. Automatic ontology construction from the literature. *Genome Informatics*, 13:201–213, 2002.

[4] L. Chrisman, P. Langley, S. Bray, and A. Pohorille. Incorporating biological knowledge into evaluation of causal regulatory hypothesis. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 128–139, Kauai, Hawaii., 2003.

[5] L. Grivell. Mining the bibliome: searching for a needle in a haystack?: new computing tools are needed to effectively scan the growing amount of scientific literature for useful information. *EMBO Reports*, 3(31):200–203, 2002.

[6] K. Korb and A. Nicholson. *Bayesian Artificial Intelligence*. Chapman and Hall/CRC, 2004.

[7] K. McGarry, S. Garfield, and N. Morris. Recent trends in knowledge and data integration for the life sciences. *Expert Systems: the Journal of Knowledge Engineering*, 23(5):337–348, 2006.

[8] K. McGarry, N. Morris, and A. Freitas. The integration of heterogeneous biological data using bayesian networks. In *AI-2006: the twenty-sixth Annual International Conference of the British Computer Society's Specialist Group on Artificial Intelligence (SGAI)*, pages 44–58, 2006.