Modular Multi-Relational Framework for Gene Group Function Prediction

Beatriz García Jiménez, Agapito Ledezma, and Araceli Sanchis

Universidad Carlos III de Madrid
Av.Universidad 30, 28911, Leganés, Madrid, SPAIN
beatrizg@inf.uc3m.es

Abstract. Determining the functions of genes is essential for understanding how the metabolisms work, and for trying to solve their malfunctions. Genes usually work in groups rather than isolated, so functions should be assigned to gene groups and not to individual genes. Moreover, the genetic knowledge has many relations and is very frequently changeable. Thus, a propositional ad-hoc approach is not appropriate to deal with the gene group function prediction domain. We propose the Modular Multi-Relational Framework (MMRF), which faces the problem from a relational and flexible point of view. The MMRF consists of several modules covering all involved domain tasks (grouping, representing and learning using computational prediction techniques). A specific application is described, including a relational representation language, where each module of MMRF is individually instantiated and refined for obtaining a prediction under specific given conditions.

Keywords: Multi Relational Data Mining, Gene Function, Multi-Label Relational Decision Tree, Inductive Logic Programming, Structure Data.

1 Introduction

One of the main challenges in molecular biology is gene function annotation. This task determines the function of the genes and their products, such as proteins. Bio-scientists need this kind of information in order to understand how metabolisms work, to find the reasons for specific molecule behaviours, and to design treatment solving malfunctions in some biological processes.

Gene function annotation is hard task for bio-scientists due to several reasons. First, a frequently changeable environment, caused by the improvement in the high-throughput experimental technologies that produces a huge quantity of data that is constantly renewed. Next, a suitable function annotation depends on some relevant domain knowledge, which only resided in bio-experts, not in databases. Besides, uncertainty and unknown information in biology leads us to work with a limited domain knowledge.

Last but not least, one gene does not carry out its functions alone, but the genes work in groups. These are characterised by to be genes with a similar expression profile in DNA arrays, proteins in the same complex, elements in an

Presented at the 19th International Conference on Inductive Logic Programming (ILP'09).

interaction network or with certain level of sequence similarity, and so on. Thus, the function annotation problem must be considered preferably as shared by genes in a group, instead of individual genes.

Furthermore, due to the experimental techniques are costly in resources and time, the function prediction methods have shown an interesting alternative in the last years. Besides, the high quantity of data requires the use of computational prediction methods.

To summarize, gene group function annotation is an open problem, that have not been solved currently yet. Besides, particular and very specific systems are not good solutions, due to the variability in the domain context.

This paper proposes a modular framework which address the gene group function prediction problem from a relational database point of view. The framework comprises several modules, adapting to any different application, in a flexible way.

Some biological approaches to solve the gene group function prediction problem have been developed [2, 6], but they have not taken into account the advantages of Multi-Relational Data Mining (MRDM). On the other hand, other similar biological domains have been faced with relational techniques successfully [5, 3].

This paper is organised as follows: Section 2 presents the relational approach for this domain. Section 3 explains the Modular Multi-Relational Framework. An application of the previous method is described in Section 4. Finally, in Section 5, conclusions and future work are summarized.

2 Relational Approach

In biological domain, there are many relational information, due to the intrinsic structure of the molecules and the importance of the similarity among different species (i.e. homology associations). Even more in function annotation domain, where the relations among different genes in the same group are fundamental for explaining why they work together. For instance, in a low level, molecules (nodes) are bound by chemical links (as a graph structure). Also, in a higher level, there are interaction networks composed of functional connections (links) among proteins (nodes). Thus, we think that MRDM is more suitable for solving this problem than traditional propositional Data Mining (DM).

Additional advantages of MRDM over the propositional DM approach are: (a)a decrease in the number of redundant features and missing values (very common facts in biological domains); (b)a better representation of real world problems, without losing the semantic after a proposionalization process; (c)an improved storage and management of the data, organised in modules or tables, according to the relations; what makes easier work with many data, different logic predicates and diverse data sources; (d)an easier representation of structured information, such as networks or graphs in interaction networks, pathways or semi-structured data from text mining results.

3 Modular Multi-Relational Framework

For dealing with gene group function prediction domain with MRDM, we have designed the Modular Multi-Relational Framework (MMRF). We have realized that in biology the expert knowledge is distributed among many places and changes along the time, so the information must be included in different not ordered steps. Hence, the framework is modular, in order to tackle these problems which need an iterative process of refinement. The modular approach covers all domain activities: grouping, representing and learning with computational prediction techniques.

As the Figure 1 shows, the MMRF is splitted into six modules. Each module consists of one or several abstract tasks, explained further. Each module and task should be individually instantiated in a particular application of the framework, as section 4 illustrates.

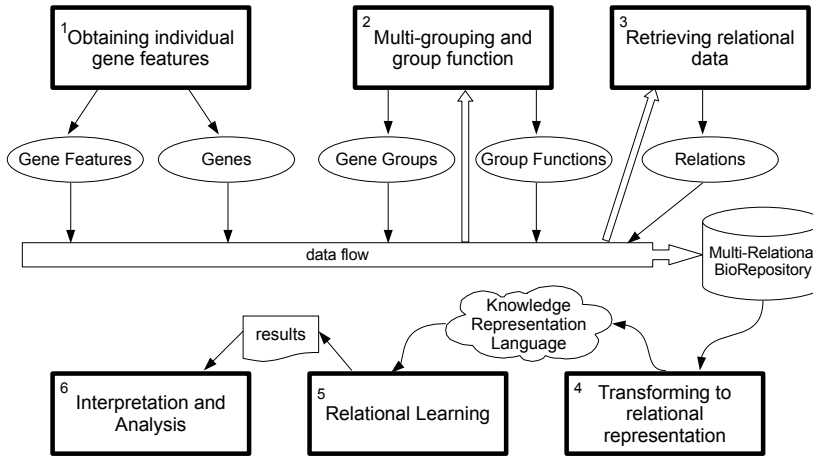

Fig. 1. An schema of the Modular Multi-Relational Framework (MMRF). The rectangles represent modules and the ellipses represent data.

1. **Obtaining individual gene features.** In this module, the tasks involved are: (a)selecting the organism, (b)choosing available and relevant gene features, and (c)searching the data source and collecting data.
2. **Multi-Grouping and group function.** This module includes two main activities: (a)making groups with the organism genes, and (b)assigning functions to each group has been created. In addition, other tasks included in this module are (c)selecting the function catalogue (i.e. the nomenclature or vocabulary used) and (d)collecting data for this particular module.
   This module is called "multi-" because multiple criteria may be defined when making groups of genes. Some of these criteria could be: genes in the same

regulation network, protein complex, pathway, or protein interaction network; genes with similar patterns in expression profiles from microarrays; with the same cellular location or protein family; common phenotypical data (for instance, pathology or tissue), or a combination of several criteria. The groups are not necessarily disjointed, e.g. they can share genes.
After groups are defined, simple methods [9] are usually applied for assigning their functions (task 2.b). Particularly, given the individual functions which annotate each gene in the group, a criterion of union, intersection or hybrid among them is used. These functions will be the classes in the further supervised learning process.

3. **Retrieving relational data.** The main task is (a)selecting relational data (common features of a subset of genes). Also, a secondary task is (b)extracting these data from selected sources. The majority of the criteria for making groups mentioned in the task 2.a are also relational data.
4. **Transforming to relational representation.** Designing the relational database and defining the knowledge representation language. Then, the collected data is transformed into the relational representation suitable for learning with a multi-relational technique.
5. **Relational Learning.** In this module, a machine learning algorithm is applied in order to obtain an easy interpretable classification model.
6. **Interpretation and Analysis.** It consists of (a)a biological interpretation of the results, including prediction evaluation, and (b)an analysis in terms of computational measures (accuracy, precision and recall, ROC curves, etc.).

4 Framework Application

We are developing an application of the MMRF for gene group function prediction. This section describes the instantiation of each one of the modules, that is the particular value for each task in each module. So far, we have already developed completely modules 1 to 4, and we have defined the procedure for applying modules 5 and 6, as it is explained further. Currently, we are working in module 5, looking for a classification model according these descriptions.

1. **Obtaining individual gene features.** (a)The organism selected is yeast (*S.cerevisiae*), a simple but eucariotic species. (b)The features are extracted from gene sequences and from the corresponding proteins. Some of these features are: the gene length, the chromosome name, the gene biotype, the protein family domain, if it is or not a transmembrane domain, etc. (c)This data is retrieved from Ensembl project [7], through the BioMart tool [10].
2. **Multi-Grouping and group function.** (a)Genes are grouped by protein complexes. (b)A function shared by a 60% of the genes in a group belongs to the assigned group functions. (c)The functional catalogue chosen is Gene Ontology (GO) (in particular GOSlim:Biological Process). We have selected GO because it is the most spread and used annotation vocabulary. (d)The protein complexes (groups) are high-throughput experimental data extracted from the detailed Krogan et.al. study [8].

3. **Retrieving relational data.** (a)The relational information consists of protein-protein interaction and homology data: paralogs from relations between genes from yeast, and orthologs from relations among yeast and another species from different categories, in particular: mouse, cow, human, fugu, chimpanzee, drosophila, xenopus, hyrax and platypus [7]. All of them represent binary relations between a pair of genes. (b) The protein-protein interactions are retrieved from [8] and the homologs through BioMart [10].
4. **Transforming to relational representation.** Our relational database design matches all the data collected in the previous modules. The knowledge representation language is defined as first-order logic predicates, in a prolog syntax (see Figure 2.a).
5. **Relational Learning.** The algorithm chosen is the Top-down induction of logical decision trees, TILDE [1], implemented in the ACE tool. Before applying TILDE, we make a data pre-process inspired by other works [3, 11], in order to get a multi-class and multi-label learning, since a group of genes have not a unique function, an important point in this domain.

```
gene_in_group(groupID,geneID).
group_function(groupID,goID).
interaction_pair(protID,protID,score).
gene(geneID,chrom,length,strand,percGC).
protein(protID,geneID).
interpro_domain(geneID,interproID).
transmembrane_domain(geneID).
ortholog(geneID,geneID,identity,id2,type).
...
```

(a)

```
gene_in_group(A,G1),protein(P1,G1)?
+-yes: interaction_pair(P1,P2,S),protein(P2,G2),
|      gene_in_group(A,G2)?
| +-yes: gene_function(G1,F1),gene_function(G2,F1)?
| | +-yes: group_function(A,F1).
| | +-no: ...
| +-no: ortholog(G1,G3,ID,X,Z), ID > 70?
| +-yes: ortholog_gene_function(G3,F2)?
| | +-yes: group_function(A,F2).
...
```

(b)

Fig. 2. On the left (a), part of an example using the knowledge representation language. On the right (b), a simplified example of a possible Relational Decision Tree.

6. **Interpretation and Analysis.** (a)The biological interpretation of the Relational Decision Tree is carried out by a bio-informatic expert, visualizing the model. It has to be a human task, since the tool does not automatically generate model interpretations. Some possible interpretation examples of an hypothetical classification model could be *"The yeast protein complex 229 is predicted to work as small molecule transporter and as component to catabolic processes (according to GO:0006810 and GO:0009056 terms)."*; or *"**if** in group A the yeast gene G1, **and** G1 has as ortholog gene G3 in mouse with higher 70% identity, **and** G3 has the function F2, **then** the group A is predicted with the function F2."* (corresponding to the below branches in the decision Tree in Figure 2.b); or hypothesis with more sophisticated relations and predictions. (b)The computational analysis of the learning results is done using precision-recall measures, because in this domain the positive predictions are the most interesting ones.

5 Conclusions and Further Work

We have proposed a new approach for addresing the gene group function prediction problem: the Modular Multi-Relational Framework. This is defined as: (1) *Modular*, i.e. a flexible approach, because this biological domain entails many possible variations in the conditions of the prediction problem, and (2) *Multi-Relational*, due to the gene group data have an intrinsec relational structure. Another contribution is a relational representation of gene group function prediction domain, in form of first-order logic predicates. Besides, an application of the MMRF has been presented, about yeast genes grouped by complexes.

As future work, first, we plan to refine modules 5 and 6 for ending the application described in section 4. Second, we also plan to define new applications by specifying all the module instantiations according to the knowledge and new developed techniques both in biological (modules 1-3, 6) and in computational terms (modules 4-6), supporting by the Structural Computational Biology Group in Spanish National Cancer Research Centre (CNIO). In biological terms, we can change group criteria in module 2, the kind of relational data in module 3, or the organism to human genes in module 1. Finally, in computational terms, we can define a module 5 which includes probabilities in relational decision tree (similar to [4]) for taking advantages of Statistical Relational Learning, since biological data entails uncertainty.

Acknowledgements. The research reported here has been supported by CICYT, TRA2007-67374-C02-02 project.

References

1. Blockeel and De Raedt. Top-down induction of logical decision trees. *Artificial Intelligence*, 101 (1-2):285–297, 1998.
2. Al-Shahrour et.al. Fatigo+: a functional profiling tool for genomic data. *Nucleic acids research*, 35:91–96, 2007.
3. Blockeel et.al. Decision trees for hierarchical multilabel classification: A case study in functional genomics. volume 4213, pages 18–29. Springer Verlag, 2006.
4. Chen et.al. Protein fold discovery using stochastic logic programs. In *Probabilistic Inductive Logic Programming*, volume 4911, pages 244–262, 2008.
5. Clare et.al. Functional bioinformatics for arabidopsis thaliana. *Bioinformatics*, 22(9):1130–1136, 2006.
6. Dennis et.al. David: Database for annotation, visualization, and integrated discovery. *Genome biology*, 4(5):P3, 2003.
7. Hubbard et.al. Ensembl 2009. *Nucleic acids research*, 37:690–697, 2009.
8. Krogan et.al. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084):637–643, 2006.
9. Sharan etal. Network-based prediction of protein function. *MolSystBiol*, 3:88, 2007.
10. Smedley et.al. Biomart-biological queries made easy. *BMC Genomics*, 10:22, 2009.
11. Vens et.al. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214, 2008.