MMRF for Proteome Annotation Applied to Human Protein Disease 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. Biological processes where every gene and protein participates is an essential knowledge for designing disease treatments. Nowadays, these annotations are still unknown for many genes and proteins. Since making annotations from in-vivo experiments is costly, computational predictors are needed for different kinds of annotation such as metabolic pathway, interaction network, protein family, tissue, disease and so on. Biological data has an intrinsic relational structure, including genes and proteins, which can be grouped by many criteria. This hinders the possibility of finding good hypotheses when attribute-value representation is used. Hence, we propose the generic Modular Multi-Relational Framework (MMRF) to predict different kinds of gene and protein annotation using Relational Data Mining (RDM). The specific MMRF application to annotate human protein with diseases verifies that group knowledge (mainly protein-protein interaction pairs) improves the prediction, particularly doubling the area under the precision-recall curve.

Keywords: Relational Data Mining, Human Disease Annotation, Multi-Class Relational Decision Tree, First-Order Logic, Structured Data.

1 Introduction

Functional annotation consists of attaching biological information to gene and genetic product sequences. For instance, identifying whether a gene is involved in a biological process, a regulation network or a molecular function; or assigning to a protein its expression profile or phenotype (tissue or disease association). Knowing the processes in which genes and proteins are involved is an essential knowledge to design disease treatments.

Nowadays, a gene/protein appears annotated in multiple distributed repositories. However, many proteins have still few or no annotation in a large number of species, because experimental techniques are costly in resources and time. This process is also overwhelmed by the high amount of data that need to be acquired and managed. Therefore, computational prediction methods have shown an useful alternative in the last years [16], in order to focus the experimental verifications on the hypotheses (predicted annotations) that are more likely to be true.

Many diverse prediction techniques have been proposed to solve the genome[1] annotation problem. Each method uses different kind and amount of input data, and is focused on a particular prediction goal. This variability in methods makes difficult a comparison among them. The simplest prediction approach is based on sequence similary, as Blast2GO [1], only useful for Gene Ontology (GO) annotation. Others predictors just include sequence and structure features [13]; while more sophisticated methods integrate heterogeneous data sources, such as Fatigo [1] and DAVID [5]. Some techniques simplify the data representation to numerical features, applying subsymbolic machine learning algorithms [10,12]; but others preserve the intrinsic structure of biological data, applying Multi-Relational Data Mining (MRDM). These techniques take advantages of the interpretable symbolic representation, such as [4,7,21] in functional annotation and [19,20] in other related bioinformatic domains.

Despite all these efforts, the proteome annotation problem remains open. We do not know functions and tasks for all proteins, and many annotations are neither verified by experts nor complete in all the biological fields of interest. Particularly, there are few specialized predictors in disease annotation, being an essential knowledge to design new drugs. Morbid OMIM (Online Mendelian Inheritance in Man) [2] contains information on all known mendelian disorders and associated genes. It is the most complete and updated repository about genetic disorders. This repository is carefully curated and frequently referenced by biological and medical scientists. For these reasons, we decide to use OMIM instead of other less known disease vocabulary such as eVOC pathology [11]. Most annotation methods using OMIM perform search rather than prediction. Some approaches predict new annotation [14], but none applies MRDM.

To summarize, genome and proteome annotation prediction is still an open problem with regard to various kind of specific annotation. Disease annotation is one of special interest. This paper proposes applying MRDM to a relevant annotation domain: human disease prediction. Besides, we want to verify the relevance of biological group relations using data integrated from different data sources. This group data is very suitable to be exploited by relational learning. We address this problem adapting a generic and flexible framework, MMRF [6], which can easily predict different annotations.

Several facts support this proposal. First, MRDM have been succesfully applied to other related bioinformatic domains. Second, we use up to date data from different biological databases used in many current science projects. Finally, we have made a special effort in data collection, selecting only experimental data, when it is possible, in order to avoid indirect redundancies coming from internal predictions from other applications, which can bias the results.

This paper is organized as follows: Section 2 briefly explains the Modular Multi-Relational Framework. The human protein OMIM disease prediction domain is described in Section 3. Section 4 presents and analyzes the application results. Finally, in Section 5, conclusions and future work are summarized.

[1] In annotation context, the terms *gene* and *protein* or *genome* and *proteome* are indistinctly used.

2 Modular Multi-Relational Framework

Modular Multi-Relational Framework (MMRF) [6] is a system originally designed for gene *Group* function prediction domain, facing the problem from a relational and flexible point of view. It has been applied to predict function for *S.cerevisiae* (i.e.Yeast) genes grouped by complexes [7]. Now, we adapt the framework since we have realized that group annotation problem is very complex to face in a single step [7]. The changes aim to solve gene and protein *individual* function annotation prediction problem, instead of *group* annotation. Nevertheless, this MMRF layout can also be considered the first phase for the group annotation problem. The complete process could be achieved obtaining first annotations for individual group elements using MMRF, and then combining them for group annotations using an alternative method (for example, union or intersection of individual annotations).

MMRF preserves the same main properties as the original layout. It is designed by modules for managing the high variability that the functional annotation biological domain entails. This facilitates changing independiently data, criteria and methodology. MMRF uses a multi-relational approach (in representation and learning) for fitting the intrinsic relational structure of gene and protein group data, and for integrating different data sources.

Figure 1 shows the new MMRF layout oriented to individual protein annotation prediction. Module 2 is now called *Selecting annotation* where the annotation vocabulary is chosen and assigned to individual gene or protein. The relational knowledge about belonging to specific biological groups (i.e.metabolic pathways, regulation networks, etc.[6,7]) is handled in module 3.

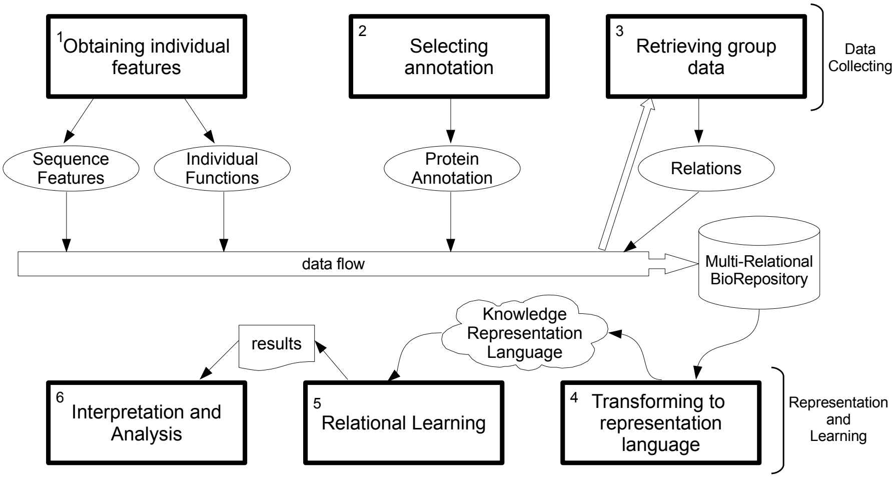
1 Obtaining individual features
2 Selecting annotation
3 Retrieving group data
Data Collecting
Sequence Features
Individual Functions
Protein Annotation
Relations
data flow
Multi-Relational BioRepository
Knowledge Representation Language
results
6 Interpretation and Analysis
5 Relational Learning
4 Transforming to representation language
Representation and Learning

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

3 MMRF Applied to Human Protein Disease Prediction

This section describes the MMRF module instantiations for applying the framework to the Human Protein OMIM Disease Prediction domain.

1.Obtaining individual features. In this application, there are 7 features derived from gene sequence (such as chromosome name and length, or transcript count) and 27 features from protein sequence (including length, positive and negative charge from the sequence, aminoacid composition and whether the protein contains a transmembrane, signal or coiled-coil domain). Also, 5 different kinds of protein functional annotations are collected, related with protein family (from *Pfam*), protein domain (from *InterPro*), biological process, cellular component and molecular function. The last three come from *Gene Ontology (GO)*, and are only from experimental results, ignoring automated annotation, for avoiding biases induced by overlaps with others annotation sources.

The numerical protein sequence features are computed with BioWeka [8] using as input the *UniProt* aminoacid sequences in FASTA format. Only *Swiss-Prot* sequences are included, because the remainder (*TrEMBL* sequences) have not been reviewed by experts. The rest of features are retrieved from *Ensembl* project, through the BioMart tool [17]. See the module 1 instantiation schema in Supplementary Material.

2.Selecting annotation. The annotation goal is genetic disorders using gene-disease associations from Morbid OMIM [2]. We apply a manual OMIM disorder categorization made by experts [9]. These disease categories have been recently used in other studies [14]. Thus, the 4,927 OMIM disorders [2] are categorized in 23 disease classes based on the affected physiological system. Some of the classes are: neurological, cancer, cardiovascular, inmunological or endocrine disease (see a complete list in Supplementary Material). Therefore, this MMRF application classifies proteins in these general disease categories [9]. However, a simple modification in MMRF module 2 could easily build a particular predictor for diseases at lower level, for instance, knowing in which specific kind of cancer (leukemia, melanoma, breast cancer, etc.) a protein is involved.

3.Retrieving group data. Two sources of group data are included, though it could be easily increased with others, as protein complexes or co-expresion data. The first data source consist of protein-protein interaction pairs, retrieved from BioGRID repository (2.0.59 Release) [18], which integrates important interaction databases as MINT, IntAct or HPRD. We select BioGRID pairs from real binary relations identified by evidences codes *Co-crystal structure, Far Western, FRET, PCA* and *Two-Hybrid*. These interactions do not include pairs split off from N-ary relations. It results in 21,687 proteins with 229,407 interactions among them. The second data source comprises metabolic pathways, which correspond to the 52 top-level human Reactome [15] pathways including 5,128 proteins, on average 159.85 proteins per pathway. See the module 3 instantiation schema in Supplementary Material.

[2] From OMIM Morbid Map on November 17th, 2009.

Data sources collected in modules 1 to 3 use different gene or protein identifiers. The original identifiers are all mapped to Ensembl (gene or protein) IDs using the cross-references from BioMart [17] queries.

4.Transforming to representation language. The knowledge representation language is a subset of first-order logic. All the collected data previously described is represented as predicates in Prolog syntax (see Figure 2). Since for humans, we can not assume the simplification *1-gene:1-protein*, as simpler organism does, the representation language has to handle information level with regard to gene and protein. These levels are related by the *1-gene:N-proteins* relation (represented as N binary predicates `protein_gene/2` per gene). Thus, the different features are separately associated to genes (predicates with *geneID* as key) or to proteins (predicates with *protID* as key) (see Figure 2). Moreover, the group data has a different representation depending on the number of elements in the group. Binary relations are represented as pairs (i.e. `ppinteracion_pair/2`). N-ary relations are represented with one group identifier plus N binary predicates (i.e. `protein_in_pathway/2`), where each predicate relates a group element with the group identifier.

```
gene(geneID,name,length,strand,trCount).            gene_biotype(geneID,bioType).
protein(protID,length,posCharge,negCharge).         protein_class(protID,omimID).
aa_composition(protID,aaID,proportionAA).           protein_gene(protID,geneID).
go_annotation_bioProcess(protID,goID).              transmembrane_domain(protID).
ppinteraction_pair(protID,protID).                  ncoils_domain(protID).
protein_in_pathway(groupID,protID).                 pfam_domain(protID,pfamID).
...
```

Fig. 2. Fragment of the knowledge representation language in proteome disease prediction domain

The instantiation of module ***5.Relational Learning*** consist of applying the algorithm Tilde [3], implemented in the ACE tool, using a multi-class and multi-label learning, inspired by other works [21]. The instantiation of module ***6.Interpretation and Analysis*** consist of evaluating the result with Precision-Recall curves (PRC). For more details, see a previous MMRF application [7], which shares the same instantiations of modules 5 and 6.

4 Results and Discussion

This section describes the results of predicting human protein diseases with MMRF. The whole data set comprises 6,958 protein-disease annotations, for 5,640 different proteins (examples) and 21 diseases (classes). Each protein can be associated with more than one disease, ranged from 1 to 5, in this set. On average, there are 331.3 annotations per disease. There are at least 40 proteins per class (the two classes with less than this minimum have been ignored). On average, there are around 5% positive vs 95% negative examples per class, although the protein class distribution is not equitable (see Supplementary Material). Each of the four majority classes has more than 10% of all annotations.

The background knowledge also includes related proteins without disease associations, but belonging to a metabolic pathway or having an interaction with a disease protein.

We compare two configurations, which differ in module 3 instantiation, it means on group relational data used for learning. The first one (***a.- Without groups***) does not included neither pathway nor protein-protein interaction data. The second configuration (***b.- With groups***) includes both kind of data from biological groups. In addition, we analize the learning implications of relational knowledge representation for groups.

The results shown in Table 1 and Figure 3 come from three folds cross validation experiments. Table 1 shows several quantitative measures and Precision-Recall curves appear in Figure 3 for the two configurations. All of them are the average results about overall 21 classes.

Table 1. Quantitative results from human protein disease prediction with MMRF. AU(PRC): Area Under Precision-Recall Curve. MSE: Mean Squared Error.

	Relational knowledge	
	a) without groups	**b) with groups**
AU(PRC):	0.282	0.625
Correlation:	0.290	0.599
MSE:	0.049	0.034

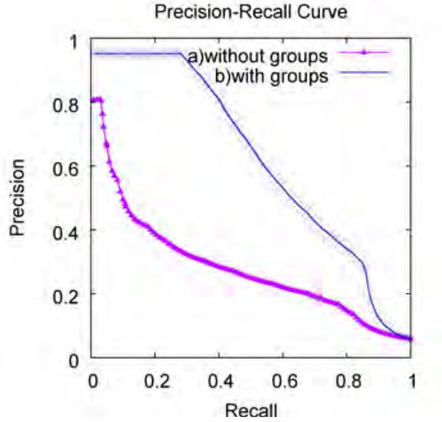

Fig. 3. Precision-Recall curves from human protein disease prediction with MMRF, in different configurations

Table 1 and Figure 3 point out that prediction with group data (configuration “b”, on the right) improves the double upon without groups (configuration “a”, on the left), in both AU(PRC) and correlation. Hence, this comparison asserts that knowledge about proteins belonging to a biological group is very relevant in disease annotation prediction.

Figure 4 presents a fragment of a relational decision tree from configuration with group data (b). The first tree node (see line 3) determines that `ppinteraction_pair/2` (a protein-protein interaction relation) is the most discriminative predicate. This fact confirms the relevance of group knowledge to predict annotations. Moreover, in the first ’yes’-branch (line 4), the second node includes a typical feature in protein function prediction: the positive charge of protein sequence [12] (variable Y), partially supporting the model reliability. Besides, in the first ’no’-branch (line 11), the most relevant query includes a *N:1* relation (predicate `protein_gene/2` relates a protein with the gene it comes from), emphasizing the high influence of relational knowledge on the prediction.

```
class(-A,-B,-C,-D,-E,-F,-G,-H,-I,-J,-K,-L,-M,-N,-O,-P,-Q,-R,-S,-T,-U,-V)
[0.011436170212766] 3760.0
ppinteraction_pair(A,-W),not(W=A) ?
+--yes: [0.0188476036618201] 1857.0
|       protein(W,-X,-Y,-Z),Y>=0.107506 ?
|       +--yes: [0.0219123505976096] 1004.0
|       |       ppinteraction_pair(W,W) ?
|       |       +--yes: [0.0569948186528497] 386.0
|       |       |       transmembrane_domain(W) ?
...
+--no:  [0.00420388859695218] 1903.0
        protein_gene(A,-M26),gene(M26,-N26,-O26,-P26,-Q26),O26>=84418 ?
        +--yes: [0.00981996726677578] 611.0
        ...
```

Fig. 4. Fragment of a relation decision tree in configuration with groups

Therefore, the importance of protein-protein interaction and protein-gene relations indicates that Relational Data Mining is essential in this domain. This is because to propositionalize this kind of data would be very complex or resulting in having redundant data. For instance, for protein binary relations, the single attribute-value table should have thousands of Boolean attributes, one per each protein. In addition, it should repeat all the gene features as attributes for all proteins that come from the same gene. Furthermore, attribute-value learning can not represent knowledge or retrieve hypotheses about features of related genes and proteins, as tree fragment in Figure 4 shows.

5 Conclusions and Further Work

This work highlights the relevance of biological group data for annotation prediction, particularly in proteome disease association. Since the most efficient and viable representation of this group knowledge is with relations, relational learning and the Modular Multi-Relational Framework are confirmed as very suitable for solving the proteome annotation problem. This is particularly relevant since the data comes from the integration of multiple up to date biological databases. Besides, the hypotheses learned through Relational Data Mining are mostly unfeasible to achieve in attribute-value learning and it holds the advantage of being readable for biology experts. This work has two main differences from a previous MMRF application [7]. For the annotation goal, diseases from OMIM morbid are used instead of general functions of GO Slim. Moreover the organism has been changed from yeast to human, which is more complex but more interesting. Thus, the obtained predictor let us select a subset of unknown protein-disease association (the most likely predictions) to be verified by in-vivo experiments.

As further work, many alternatives appear. It would be interesting to make a comparison between this overall classes predictor and 21 independent predictors, one per each disease class. Other possibilities would be related to biological MMRF applications. For instance, including new group data, such as protein complexes or co-expression data; applying the predictor to annotate unknown proteins; or changing the prediction goal to a different annotation field, as predicting if a protein belongs to a metabolic pathway.

Acknowledgments. This research has been supported by CICYT, TRA 2007-67374-C02-02 project and by the expert biological knowledge of the Structural Computational Biology Group in Spanish National Cancer Research Centre (CNIO). The authors would like to thank members of TILDE tool developer group in K.U.Leuven for providing their help and many useful suggestions.

Supplementary Materials (for online version). They include two schemas about obtaining individual features and retrieving group data, the complete list of 23 disease categories and a figure showing the protein per disease distribution.

References

1. Al-Shahrour, F., et al.: Babelomics: a systems biology perspective in the functional annotation of genome-scale experiments. Nucl. Acids Res. 34, W472–W476 (2006)
2. Amberger, J., et al.: McKusick's Online Mendelian Inheritance in Man (OMIM(R)). Nucl. Acids Res. 37, D793–D796 (2009)
3. Blockeel, H., De Raedt, L.: Top-down induction of logical decision trees. Artificial Intelligence 101(1-2), 285–297 (1998)
4. Clare, A.: Machine learning and data mining for yeast functional genomics. PhD thesis, University of Wales, Aberystwyth (2003)
5. Dennis, G., et al.: DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology 4(5), P3 (2003)
6. García, B., et al.: Modular Multi-Relational Framework for Gene Group Function Prediction.. In: Online Proceedings ILP (2009)
7. García Jiménez, B., Ledezma, A., Sanchis, A.: S.cerevisiae complex function prediction with modular multi-relational framework. In: García-Pedrajas, N., Herrera, F., Fyfe, C., Benítez, J.M., Ali, M. (eds.) IEA/AIE 2010. LNCS, vol. 6098, pp. 82–91. Springer, Heidelberg (2010)
8. Gewehr, J., et al.: BioWeka extending the Weka framework for bioinformatics. Bioinformatics 23(5), 651–653 (2007)
9. Goh, K., et al.: The human disease network. PNAS 104(21), 8685–8690 (2007)
10. Jensen, J., et al.: Prediction of human protein function according to Gene Ontology categories. Bioinformatics 19(5), 635–642 (2003)
11. Kelso, J., et al.: eVOC: A Controlled Vocabulary for Unifying Gene Expression Data. Genome Research 13(6a), 1222–1230 (2003)
12. Lee, B., et al.: Identification of protein functions using a machine-learning approach based on sequence-derived properties. Proteome Science 7(1), 27 (2009)
13. Lee, D., et al.: Predicting protein function from sequence and structure. Nature reviews. Molecular Cell Biology 8(12), 995–1005 (2007)
14. Linghu, B., et al.: Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biology 10(9), R91 (2009)
15. Matthews, L., et al.: Reactome knowledgebase of human biological pathways and processes. Nucl. Acids Res. 37, D619–D622 (2009)

16. Peña-Castillo, L., et al.: A critical assessment of mus musculus gene function prediction using integrated genomic evidence. Genome Biology 9, S2 (2008)
17. Smedley, D., et al.: BioMart-biological queries made easy. BMC Genomics 10 (2009)
18. Stark, C., et al.: BioGRID: a general repository for interaction datasets. Nucl. Acids Res. 34, 535–539 (2006)
19. Trajkovski, I., et al.: Learning relational descriptions of differentially expressed gene groups. IEEE Transactions on Systems, Man, and Cybernetics 38(1), 16–25 (2008)
20. Tran, T.N., Satou, K., Ho, T.-B.: Using inductive logic programming for predicting protein-protein interactions from multiple genomic data. In: Jorge, A.M., Torgo, L., Brazdil, P.B., Camacho, R., Gama, J. (eds.) PKDD 2005. LNCS (LNAI), vol. 3721, pp. 321–330. Springer, Heidelberg (2005), http://dx.doi.org/10.1007/11564126_33
21. Vens, C., et al.: Decision trees for hierarchical multi-label classification. Machine Learning 73(2), 185–214 (2008)