Predicting Protein Relationships to Human Pathways through a Relational Learning Approach Based on Simple Sequence Features

Beatriz García-Jiménez, Tirso Pons, Araceli Sanchis, and Alfonso Valencia

Abstract—Biological pathways are important elements of systems biology and in the past decade, an increasing number of pathway databases have been set up to document the growing understanding of complex cellular processes. Although more genome-sequence data are becoming available, a large fraction of it remains functionally uncharacterized. Thus, it is important to be able to predict the mapping of poorly annotated proteins to original pathway models. **Results:** We have developed a Relational Learning-based Extension (RLE) system to investigate pathway membership through a function prediction approach that mainly relies on combinations of simple properties attributed to each protein. RLE searches for proteins with molecular similarities to specific pathway components. Using RLE, we associated 383 uncharacterized proteins to 28 pre-defined human Reactome pathways, demonstrating relative confidence after proper evaluation. Indeed, in specific cases manual inspection of the database annotations and the related literature supported the proposed classifications. Examples of possible additional components of the Electron transport system, Telomere maintenance and Integrin cell surface interactions pathways are discussed in detail. **Availability:** All the human predicted proteins in the 2009 and 2012 releases 30 and 40 of Reactome are available at http://rle.bioinfo.cnio.es.

Index Terms—Pathway relationship prediction, sequence-based prediction, knowledge relational representation, machine learning, function prediction, human reactome pathways

1 Introduction

Biomolecular pathways represent an abstract compilation of knowledge that pertains to metabolic, regulatory and signalling events, organised as cascades of protein interactions influenced by other molecules [1], [2]. Deregulation of such signalling systems has been implicated in diverse human pathologies, including cancer, neuronal degeneration, muscle atrophy, immune deficiency and diabetes [3].

Recent years have seen renewed interest in storing and annotating pathways [4], [5], although this presents several notable challenges. Among these is the growing amount of experimental information available, the obvious limitations of databases and annotation resources, and the uncertainty as to what are the boundaries of a pathway.

Due to the lack of uniformity in the definitions of pathways found in the literature [6], there may be considerable variation among the data available in different databases, such as Reactome [7], KEGG [8] and MetaCyc [9]. Indeed, efforts to develop a standardized form of annotation are currently under-way (e.g., Pathway Commons [5] and WikiPathways [10]).

Our research focuses on predicting the pathway membership of uncharacterized proteins not included in the original model pathway. Typically, additional proteins associated with a biological process of interest (such as regulators) are not considered part of the pathway for several reasons [11], due to: the introduction of indirect noise in empirical procedures; a lack of data in specific databases devoted to a particular functional area or organism, and using the classical isolated entity representation; or the subjective opinion of experts who designed the pathways based on their knowledge and experience.

It is important that we establish a relationship between previous publications and the approach presented here, which is Relational Learning-based Extension (RLE). First, our approach does not use homologies and it is not based on the extrapolation of pathways between species on the basis of sequence similarity, distinguishing it essentially from other published methods [12], [13]. Second, our approach attempts to predict pathway membership of new potentially related proteins and not to define new pathways. In this sense it is very different from the methods that analyze protein interaction networks and other features to discover new pathways [14], [15], [16]. Third, our approach differs but is related to another approach that also uses molecular interaction data to propose candidates that are part of known pathways [17]. The difference is that this earlier method [17] exclusively uses protein interactions and in the approach presented here, interactions represents only a

- *B. García-Jiménez and A. Sanchis are with the Computer Science Department, Universidad Carlos III de Madrid, Av. Universidad 30, 28911 Leganés, Madrid, Spain. E-mail: {beatrizg, masm}@inf.uc3m.es.*
- *T. Pons and A. Valencia are with the Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Melchor Fernández Almagro 3, 28029 Madrid, Spain. E-mail: {tpons, avalencia}@cnio.es.*

Manuscript received 21 Dec. 2012; revised 2 Apr. 2014; accepted 3 Apr. 2014. Date of publication 17 Apr. 2014; date of current version 4 Aug. 2014.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2014.2318730

minor component (i.e., RLE uses the similarity of features with interacting proteins instead of explicit information on interactions).

Finally, RLE uses simple sequence features as its input, which makes it part of a range of function prediction strategies that combine simple properties in different scenarios [18], [19], [20]. Of these, the most similar to RLE is the ProtFun method developed by Soren Brunak's group when it assigns different Gene Ontology (GO) Biological Process categories to human genes [19]. Although RLE uses simpler sequence properties than ProtFun, both systems use sequence features to predict an association with biological processes. Despite the apparent contradiction, these systems do not search for characteristics common to all the proteins in the pathway but rather, they look for specific characteristics associated with some of the components in the pathway. Hence, the proposed protein memberships bear similarities to some of the original pathway proteins rather than to some characteristics common to all of them. Thus, RLE looks for unannotated proteins with similar features to pathway proteins instead of physical links to pathway proteins.

One of the key differences from previous approaches is the use of a relational representation that allows individual and pair-wise information to be combined. In this case, RLE uses information on features associated to each individual sequence, as well as information about the features of neighbouring proteins in the interaction network.

The relational representation allows RLE to apply a sophisticated combination of relational and propositional machine learning algorithms to retrieve frequent patterns and to induce relational decision trees. This machine learning combination has previously been applied successfully to other functional annotation problems, such as assigning GO and MIPS terms to *Arabidopsis thaliana* [21] and *S.cerevisiae* [22] genomes, although using homology, secondary structure, sequence and expression data. Although these applications are used for predictive analyses in simpler species, they require more complex homology data than in the present study. Nevertheless, the results of these studies indicated that the machine learning combination used by RLE is readily applicable to other function prediction tasks, including those involving the sharing of common data.

Here, we defined and used the RLE system to predict putative pathway membership for uncharacterized human proteins according to the definitions of Human Reactome pathways [7], an expert-authored, peer-reviewed, manually curated pathway database widely used in biological and computational studies (see [23], [24]). The results of our predicted Reactome annotations are presented and some relevant biological cases are discussed in detail.

2 Materials and Methods

2.1 Relational Representation

Protein-protein interactions and protein complexes represent important relationships in biological pathways. For that reason, we have included these interactions in the learning process as relational information that may influence the final predictions. The classical data mining approach represents data in a propositional manner, i.e., one table featuring one row per protein and a list of columns (or features) for each specific protein. Propositional representation of the data used in the present study would require thousands of Boolean attributes per protein (one for each of the potential interacting partners in the entire proteome), and where most of the columns would have no values. By contrast, using a relational representation [25] it is sufficient to define one binary predicate and to include as many instances as true interaction partners exist. Relational representation also allows us to consider sequence features of the interaction partner in the learning process through a link with its identifier. For example, we can annotate a *protein A* with the *membrane trafficking* pathway because *protein A* is involved in a complex interaction with *protein B*, which contains a transmembrane region.

The main relational representation language is logic programming, a subset of first-order logic (also known as predicate logic) in which each element is a logical predicate. All the data collected (described in Section 2.2) are represented as logical facts in Prolog syntax (see Fig. 1 for a fragment and Text S1 and S2 in the Supplementary Material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2014.2318730, for the complete relational representation). This representation enables relational learning to be applied.

2.2 Input Data Sources

In constructing the system used here, we took data from multiple sources to build our own data set. Using amino acid sequences as the input, the three numerical protein sequence features (length, positive and negative charge) were computed with BioWeka software [26] and made discrete to increase their expressiveness (see details about discretization in Text S1 of the Supplementary Material, available online). Simple predictions or annotations from protein sequences were also included (e.g., whether the protein contains a transmembrane region, signal peptide or coiled-coil domain). The gene features used were chromosome name, length, strand and number of transcripts or isoforms (see `gene` predicate in Fig. 1). These properties were retrieved from Ensembl [27], release 56 (through to BioMart Central Portal [28]).

Two types of relationships between proteins were considered: protein-protein interactions; and protein complexes, represented in the form of interaction partners (`ppinteraction_pair` and `complex_interaction` predicates, respectively, as represented in Fig. 1). Due to the high quantity and quality of the interactions extracted from literature and high-throughput experiments, the data pertaining to protein-protein interaction pairs was retrieved from the BioGRID repository (2.0.59 release) [29]. These data are curated together with expert partners such as MINT [30], IntAct [31] and HPRD [32]. We selected BioGRID pairs from real binary relationships identified through evidence codes *Co-crystal structure, Far Western blot, FRET, PCA* and *Two-Hybrid* studies. Protein complexes were considered as protein pairs, since the databases represent complexes as pairs and the information available to rebuild the complex is neither complete nor curated. We retrieved complexes from the same BioGRID release, selecting relationships identified by evidence codes *Affinity Capture, Co-purification* and *Reconstituted complex*. We also included sequence features for the interaction partners.

```
protein(protID,length,positiveCharge,negativeCharge).
transmembrane_domain(protID).
ncoils_domain(protID).
signal_domain(protID).
protein_gene(protID,geneID).
gene(geneID,chrName,length,strand,numTranscriptsOrIsoforms).
ppinteraction_pair(protID,protID).
complex_interaction(protID,protID).
protein_class(protID,reactomeID).
```

A. Fragment of the knowledge representation language

```
protein('ENSP00000368547',380,0.107894,0.189474).
ncoils_domain('ENSP00000368547').
protein_gene('ENSP00000368547','ENSG00000171055').
gene('ENSG00000171055','2',97723,-1,15).
ppinteraction_pair('ENSP00000368547','ENSP00000204549').
ppinteraction_pair('ENSP00000368547','ENSP00000216267').
ppinteraction_pair(...).
complex_interaction('ENSP00000367830','ENSP00000368547').
complex_interaction('ENSP00000368547','ENSP00000383712').
complex_interaction(...).
```

B. Protein FEZ2_HUMAN represented by logical facts

Fig. 1. Knowledge representation language in the pathway prediction domain. `protein/4` predicate represents properties associated to a protein; `transmembrane_domain` represents a protein with transmembrane region/s; `ncoils_domain` represents a protein with coiled-coil domains/s; `signal_domain` represents a protein with signal peptide sequence; `protein_gene` represents the relation between a protein and the gene encoding it; `gene` predicate represents the properties associated to a gene; `ppinteraction_pair` represents a protein related to other protein by Protein-Protein Interaction data; `complex_interaction` represents a protein related to other protein with a described interaction in a complex and `protein_class` represents a protein belonging to a specific Reactome pathway. The goal of RLE, in knowledge representation terms, is to associate the predicate `protein_class` to those proteins without annotations in Reactome. Panel A shows main predicates in the knowledge representation language defined to this domain. In panel B, you can see an example of the set of logical facts that represent the human protein FEZ2_HUMAN (ENSP00000368547, fasciculation and elongation protein zeta 2) according to the knowledge representation described in panel A. These facts imply that features for this protein are: protein with coiled-coil domain/s; long gene sequence, since third argument (97,723) is higher than 30,447 (see Text S1 for a detailed list of thresholds); high number of transcripts, because 15 is greater than 4; and with relations in protein-protein interactions and protein complexes. `ppinteraction_pair(...)` and `complex_interaction(...)` mean this protein has more relations than those shown here, by Protein-Protein Interaction and in protein complexes, respectively.

Finally, since Reactome pathways [7] were the annotation objective, we analyzed 37 of the 52 top-level human Reactome pathways, release 30. These 37 pathways corresponded to pathways fulfilling a minimum size requirement of at least 50 proteins in the original pathway. This minimum number of proteins required by the RLE method to learn, was arbitrarily selected as a rule of thumb to avoid some bias information from scarcity data.

Briefly, we collected 22,304 genes, 72,731 protein isoforms, 229,407 protein interaction pairs, 478,420 complex interaction pairs and 37 pathways with an average of 142 non-redundant proteins per pathway.

As the data sources use different gene or protein identifiers, the original identifiers were all mapped to Ensembl (gene or protein) IDs using the cross-reference system from BioMart [28].

2.3 Building Data Sets: Training, Testing and Application

2.3.1 Training and Test Data Sets

As our goal is to capture 'functional' sequence features as opposed to sequence similarities, we built a non-redundant data set. In this way, we avoid biases in the learning process due to indirect relationships between similar proteins in the training and test data sets. The reduction in redundancy is a typical conservative process, which is the best option when the relationship between evolutionary origin and sequence features may not be readily determined.

Redundancy was removed from two terms: isoforms and sequence similarities. A single gene can express itself as several proteins or transcripts, called isoforms, produced by processes such as alternative splicing. The number of isoforms is preserved as a sequence feature for the learning process (see Fig. 1). However, to reduce the redundancy in our data set we only selected a reference isoform for all the proteins expressed by the same gene. We considered the reference isoform as the protein with most annotations in Reactome, as this is the prediction goal. In cases where several isoforms had the same number of annotations, the longest sequence was considered as the reference isoform and accordingly, we extracted 3,510 reference isoforms from Reactome. Note that 97.5 percent of our reference isoforms corresponded to the longest isoforms.

Next, a sequence-similarity reduction of isoforms was applied to the protein sequences based on BLAST [33] alignments. We applied one of the Hobohm algorithms [34] for homology reduction that had been expanded [35] and applied in previous studies [19], [36], [37]. Since RLE predictions are entirely based on amino acid sequence instead of protein structure data, the Hobohm algorithm 2 was slightly modified to define similarity based on amino acid sequences rather than protein structures. The original algorithm is based on the same dynamic sequence alignment algorithm [38] used in BLAST, and we use BLAST results as a measure of similarity.

We calculated sequence similarity in the complete human proteome with BLASTP [39] (see Section 2.2), running BLASTP with default parameters except for an E-value $= 0.01$. An E-value of 0.01 means that we expect one random match in every hundred for the given score. Setting a low threshold for the E-value (BLASTP default is 10) would reduce the number of potential errors. A sequence identity threshold of 30 percent was applied.

Following sequence similarity reduction of the main isoform proteins, we were left with 1,654 unique annotated proteins (2,762 increased by different pathway annotations) in the 37 selected Reactome pathways. Two thirds of these proteins (1,108 proteins) were randomly grouped into the training data set and one third (546 proteins) were grouped into the test data set.

One alternative validation to our two thirds for training and one third for testing would be a Cross-Validation (CV) process. However, a 10-fold CV experiment is not suitable for this specific biological domain. The main reason is the multi-class and multi-label nature of this pathway annotation problem, together with the small number of available proteins in several classes or pathways. Many pathways are small and the process of redundancy reduction sequence-similarity proteins and isoforms, make them even smaller. As explained above, redundancy reduction is essential in

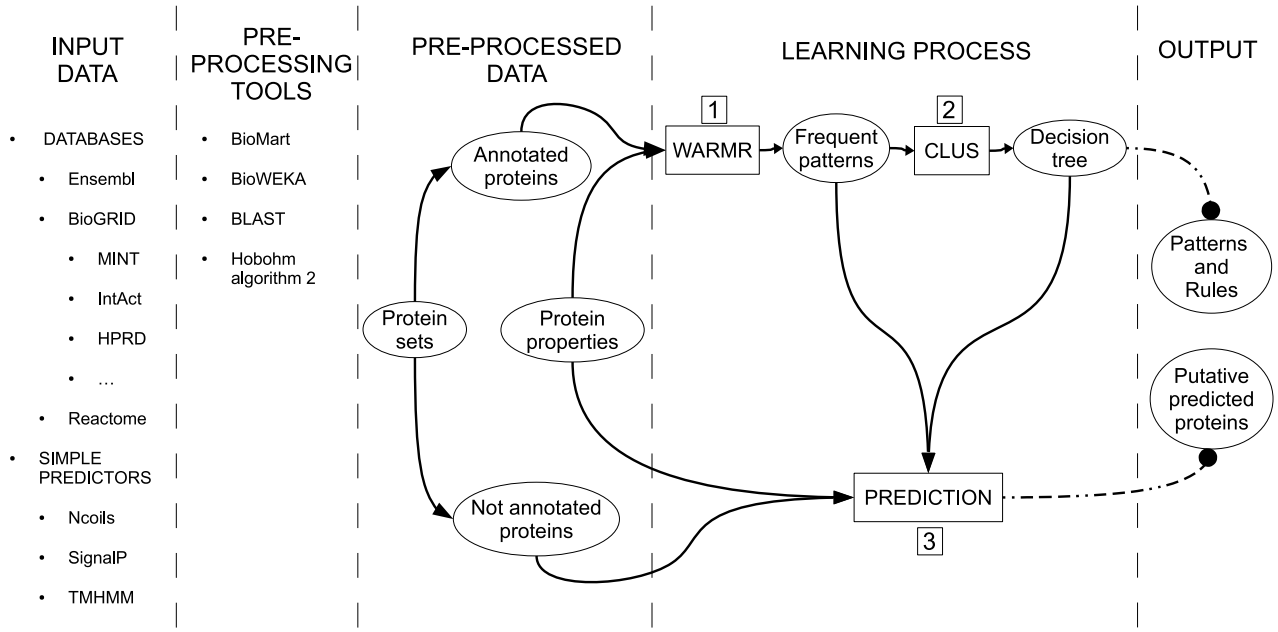

Fig. 2. Schema of the RLE system. In the learning process, the rectangles correspond to methods and the ellipses correspond to data with different knowledge representations. In the pre-processing tools, BLAST and Hobohm algorithm 2 are applied to redundancy reduction.

order to avoid biases in the learning process introduced by indirect relationships. Moreover, according to the literature, it has been suggested that stratified 10-fold CV experiments can be suitable for evaluation of solutions in imbalanced multi-class problems if the partitions can have the same composition than the full data set [40]. In our case this means that each of the 10 partitions in which the proteins are split must follow the same distribution of classes as the whole set, with a number of proteins belonging to each pathway exactly equal in each partition. With few proteins per pathway and each protein often assigned to more than one pathway (multi-label condition) makes impossible to create 10 partitions with the same proportion of proteins per pathways and therefore it is impossible to perform a non-biased CV.

2.3.2 *Application Data Set*

The proteins not annotated in Reactome (i.e., the application data set) are used as the input to identify additional proteins related to the Reactome pathways with the RLE system. Of the proteins in the 37 Reactome pathways of interest, 18,794 were not annotated in Reactome (22,304 main isoform proteins minus 3,510 with Reactome annotations). The sequence similarity reduction procedure allocated 8,187 proteins to the application data set, which was itself non-redundant as well as being non-redundant in relation to the training and test data sets. These 8,187 proteins not annotated in Reactome represent the application data set.

2.4 Prediction Method: Frequent Patterns and Decision Tree

The RLE method is split into three steps: the retrieval of relational frequent patterns; the generation of a propositional decision tree; and application of this decision tree to unannotated proteins (see Fig. 2).

In the first step, we retrieved the frequent patterns (i.e., a relevant sequence of logical facts) with the first-order logic association rule mining algorithm WARMR [41], which uses a relational data set as the input and that is implemented with the ACE tool [42]. The WARMR algorithm identifies all patterns that satisfy a language bias in the training data set and that exceed a minimum frequency. Moreover, like the propositional APRIORI algorithm [43], the WARMR algorithm performs a level-wise search that is quick and efficient in large databases. We applied WARMR to the proteins in each independent pathway to retrieve the frequent patterns that characterize each particular pathway. This is a novel application that differs from previous use of frequent patterns in combination with decision trees [21], [22] where patterns are retrieved for all examples together, without splitting them according to the different pathways or classes. Subsequently, frequent patterns for all the pathways were kept individually to construct the predictor system. Finally, frequent patterns were used as the input for the next step to generate a propositional decision tree, transforming each pattern into a Boolean attribute in function of whether the pattern was satisfied by the specific protein.

In the second step, we built decision trees with our defined training data set using the CLUS system [44]. This implements the predictive clustering tree framework that induces propositional decision trees using an algorithm similar to C4.5 [45], but that views the decision trees as hierarchies of clusters. The root-node is a cluster with all instances, which is then split into smaller clusters recursively, thereby minimizing the intra-cluster variation. This framework allows us to tackle more complex prediction problems. We selected CLUS over other decision tree algorithms as CLUS readily facilitates multi-class and multi-label learning. It corresponds to our pathway annotation problem, as the number of possible pathways is greater than two (multi-class) and each protein may belong to more than one pathway (multi-label). The relational decision tree/s obtained after applying both WARMR and CLUS to our relation data allow us to associate new similar proteins to the Reactome

pathways (Supplementary Text S2, available online, shows a further description about the representation of the results).

In the third step, we predict proteins that may be related to the Reactome pathways by applying our approach to the unannotated proteins (i.e., the application data set - see Section 2.3.2). This last step is completely new and was designed for this study, making it distinct from other earlier combinations of WARMR and CLUS. RLE predicts proteins with similar properties to the pathway, without providing explicit information about how they interact in the pathway. Our relational learning system associates a list of scores to each protein it classifies and as a result, each protein has one score associated to each pathway. Hence, we must select a list of thresholds to separate the proteins predicted to be similar to given pathway proteins from those that are not. Among the multiple options, we selected the next combination as a reasonable criterion. We sought to predict similar proteins by up to 20 percent of the non-redundant pathway size. For each predicted pathway, we sorted the proteins in the application data set by decreasing the score value and we then selected all the proteins up to that where the last change in score values reached the first 20 percent of the pathway size. In cases where no change occurred, the system did not propose any protein associated to the pathway, and the pathway threshold was the lowest score of the selected proteins. Where possible, and as an additional criterion, the system should apply more than one rule per pathway in the application data set (i.e., different branches from root to leaf) in order to maximize protein diversity.

The prediction method described can generate many different annotation systems depending on the configuration parameters used. For the current problem of predicting pathway associations, we have configured a RLE with a minimum frequency of 0.2 and a maximum depth of 4 as the WARMR parameters. These parameters were selected to build decision trees with different attributes depending on the specific pathway, since we are interested in predicting a range of proteins related to each pathway in parallel to the molecular variability of the proteins in the corresponding pathway. Of the various configurations produced, we applied a tradeoff between performance (measured as AUPRC in test data set) and diversity of rules for each pathway when making our selection. This means that a system with better performance could have been selected at the expense of decreasing (or even removing) the diversity among the proteins predicted, which would be similar to the features of just one pathway protein rather than several, as we obtain with the configuration selected (a detailed configuration for the selected system appears in Text S3 in Supplementary Material, available online).

Fig. 2 illustrates the entire workflow of the RLE system applied to annotate proteins through Reactome pathways as described in Section 2.

3 Results

3.1 Prediction Performance

The following section discusses the performance of our RLE system, both overall and in relation to each independent pathway.

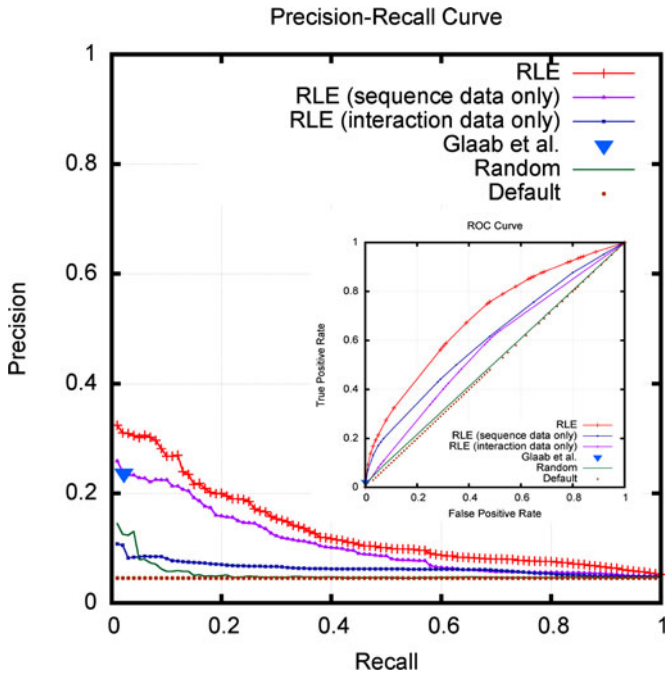

Fig. 3. Precision-Recall and ROC curves. Macro-average PR and ROC curves for the RLE system, the random and default classifiers. The default classifier corresponds to a unique leaf decision tree, which has class frequencies as scores for any given protein. The random classifier simply consists of generating a random score for each protein in the test set. The '*RLE (interaction data only)*' classifier is built according to the RLE specification with interaction data alone and the '*RLE (sequence data only)*' with the remainder. The Glaab et al. point results from applying this method to the same training and test data set as the RLE system (see Section 3.6).

We measured the performance of the test data set with Precision-Recall (PR) curves, as these curves fit our highly-skewed class distribution and addressed our interest in positive predictions in this domain (the prediction of proteins with similar characteristics to pathways rather than those without them) [46]. Moreover, standard Receiver Operating Characteristic (ROC) curves were also generated (Fig. 3).

To combine the 37 pathway-wise performance measures in an overall measure, we chose macro-average (the average area under the PR curves) rather than micro-average (the area under the average PR curve) [47]. Macro-average *Area Under Curve* (AUC) does not bias the result towards more frequent classes, providing a more homogeneous view of the results [22]. The macro-average measure computes individual curves first, one for each independent pathway, and then it averages these curves to compute a single average curve. By contrast, the micro-average measure computes a global contingency table using the sum of the scores for all pathways (true and false positives and negatives), and it then computes a single curve based on these global scores.

In quantitative terms, RLE resulted in an *Area Under the PR Curve* (AUPRC) value of 0.1337 and an *Area Under the ROC* (AUROC) value of 0.6914. While these AUPRC and AUROC values are small, they are clearly above the random and default classifiers. Combining sequence features and protein interactions, as RLE does, is better than using either sequence data or interaction data alone (Fig. 3), as evident when the RLE system was tested with subsets of input features.

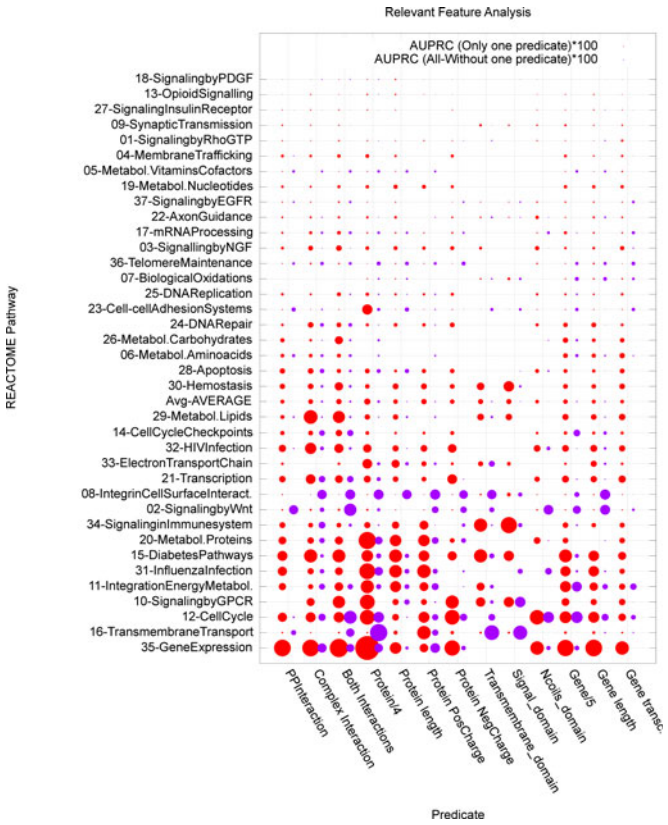

Fig. 4. Relevant learning predicate analysis. Red circles (on the left) represent features that are relevant alone and the purple circles (on the right) represent features that are relevant in combination with others. The pathways (rows) are sorted from those with the lowest AUPRC in our RLE system to those with the highest (from top to bottom).

Analyzing the performance per pathway revealed variations that fell both above and below the average values (see Table S1 in Supplementary Material, available online, for the detailed performance per pathway values). While 16 pathways exhibited greater than average AUPRC values (very reliable pathway predictions), only three pathways had AUPRC values lower than the random or default classification values (poorly reliable pathway predictions).

In general, the larger the pathway the higher the AUPRC value, with some exceptions, such as the *integrin cell surface interactions* (PathwayID:8) and *transmembrane transport of small molecules* (PathwayID:16) pathways, which represent small pathways with higher than average AUPRC values.

3.2 Relevant Feature Analysis

In the following section, we will discuss the relevant features of each pathway in the learning process.

We applied the same two relevance measures used in the ProtFun prediction method [19]. The first involves evaluating performance after training the system using each of the predicates individually (Fig. 4, see red circles representing these AUPRC values obtained in this way). The second measure shows the decrease in the AUPRC following the removal of a particular predicate. The performance of system training without one predicate is subtracted from the original AUPRC, which corresponds to the combination of all the predicates (in Fig. 4 these AUPRC differences are visualized as purple circles). The first of these measures only indicates the relevance of a given predicate by itself, without taking into account those that may also be relevant when combined with other predicates (as demonstrated by the second measure). As both measures are complementary, a predicate is deemed relevant if either of these measures is high [19].

With our relational data representation, logical predicates rather than features were used as input data (see predicates and their arguments in Fig. 1). Accordingly, the columns in Fig. 4 correspond to predicates, single arguments or an aggregate of several predicates.

By analyzing the graph (Fig. 4) it was evident that for the *'average'* pathway (in the middle of Fig. 4), no predicate is more important than others and thus, no feature contributes disproportionately. Moreover, the relevance is clearer as the prediction improves. Obviously no predicate is relevant in cases of poor prediction (upper part of Fig. 4) and for example, while the `transmembrane_domain` would be expected to be relevant in the *membrane trafficking* pathway (PathwayID:04), no significant relationship was detected. In cases of reliable predictions (below the *'average'* pathway in Fig. 4), we found clear differences in the relevance of distinct predicates within the same pathway. The most important predicate (i.e., the column with the largest points) was `protein/4`, which is an aggregate of the most discriminating features of proteins: `protein length` and `positive charge`. By contrast, interactions were not such fundamental features for the learning process (first three columns).

When some specific cases were analyzed, almost all isolated predicates performed well for the *gene expression* pathway [35], while all predicates were dependent on one another for *transmembrane transport* [16], *integrin cell surface interactions* [08] and *Signaling by Wnt* [02], with little independent contribution (several purple circles of similar size).

3.3 Prediction of Putative Proteins Related to Reactome Pathways

We check the capacity of our RLE system to identify additional proteins related to curated Reactome pathways when applied to unannotated proteins.

Following the procedure described above (Section 2.4), RLE associated 383 uncharacterized proteins (including 329 distinct proteins) to 28 pathways. The 37 original nonredundant pathways consist of 2,762 proteins, of which 1,654 are distinct proteins. In terms of the importance of the diversity in the rules that predict proteins in the same pathway, note that RLE applies several rules in 15 pathways. Therefore, our system attributed strong molecular variability to the proteins related to the same pathway in more than half of the predicted pathways. A detailed summary with respect to class is provided in Table S1 of Supplementary Material, available online.

RLE predicts proteins that share specific sequence features with one or more of the proteins in the original pathway but not with the overall characteristics of the pathway. The proteins predicted by RLE (hereafter referred to as *predicted proteins*) have not been annotated previously in Reactome and they were not redundant to Reactome annotated proteins in terms of sequence (see Section 2.3.2). Given the biased nature of Reactome pathways, the predicted proteins also share an intrinsic bias to these RLE design considerations, evident when the frequency of the properties shared

```
IF id(A),ppinteraction_pair(A,B),not(B=A),protein_gene(B,C),
         gene(C,D,E,F,G),E<3860,protein_gene(A,H),
         gene(H,I,J,K,L),J>30447 = 1 AND
   id(A),ppinteraction_pair(A,B),not(B=A),
         complex_interaction(B,C),not(C=A),not(C=B),
         protein_gene(B,D),gene(D,E,F,G,H),F<3860 = 0 AND
   id(A),complex_interaction(A,B),not(B=A),
         signal_domain(B),ncoils_domain(B) = 1
THEN REACT_11061 (Signalling by NGF)
```

Fig. 5. Example of rule or patterns resulting from the RLE system prediction. This rule associates the human protein FEZ2_HUMAN represented with logical facts in Fig. 1 to the *Signalling by NGF* pathway (see main text for details).

by the RLE predicted proteins was compared to those of the proteins used for training and to test the system (Supplementary Fig. S3, available online). We intended to find unique proteins, although we also checked whether their homologues were also predicted. The Reactome non-annotated protein data sets (sequences redundant to our application data set), contain some homologues of the predicted proteins. RLE does not predict almost any of these homologues, because RLE learns with features other than amino-acid sequences, the only input to define homology. Since RLE uses additional properties (number of transcripts, gene length, protein-protein interactions and complexes) RLE should not predict the homologous partner if there were differences in these properties (mainly due to the discrepancy in the interaction annotations or in the sequence features of the interaction partner).

We complemented the results by searching for similar annotations that could provide an independent evaluation of our predictions. Indeed, a semantic similarity analysis showed that on the basis of GO Biological Process annotations, most of the original pathways were more semantically similar to RLE predicted proteins than to random predictions.

3.4 Representation of Resulting Patterns

In computational terms, the explicit learning output of the RLE system is represented as a set of rules defining the patterns that fulfill a particular protein to be associated to a specific Reactome pathway. Fig. 5 shows an example of rule (or patterns) defining association of proteins to the *Signalling by NGF* pathway, according to the RLE prediction. In this example (see also description of the Fig. 1 legend and Supplementary text S1 and S2, available online), the first logical conjunction of this rule represented by the three first lines, means the predicted protein FEZ2_HUMAN (letter 'A') is related to the protein 'B' by a protein-protein interaction. The corresponding gene of the interaction partner 'B' (gene 'C') has a short sequence ($E < 3{,}860$), and the corresponding gene of protein 'A' (gene 'H') has a long sequence ($J > 30{,}447$). In the next logical conjunction, one or several of the logical predicates are false (ends with "=0"). Finally, the last logical conjunction means our main protein 'A' interacts in a complex with other protein characterized by a signal peptide sequence and at least a coiled-coil domain.

Supplementary Text S2 describes a different example of rule giving more details. In addition, in the RLE web (http://rle.bioinfo.cnio.es) there are links to rules and simple sequence features for each predicted protein in the studied pathways.

3.5 Biological Interpretation of the Predicted Proteins

Using complex combinations of simple properties, as reported elsewhere [18], [19], [20], it is difficult to interpret our RLE results according to those properties through a general analysis. However, it is possible to analyze the predictions for a given pathway or protein of interest by studying the frequency of predicates. It means, to compare predicates of the predicted protein(s) with those of the annotated proteins in the pathway (see Supplementary Fig. S2, available online). If possible, also to compare the predictions using different computational methods and complementary information available in databases.

To further investigate the biological implications of the RLE prediction and confirm that prediction is consistent with patterns in the input, UniProt functional annotations and the literature available on the predicted proteins were analyzed. As examples of *de novo* predictions, we used the *Electron transport chain*, *Telomere maintenance* and *Integrin cell surface interactions* pathways, due to the tendency towards a similarity between more frequent predicates and UniProt annotations, as it is explained in the next paragraphs.

In the *Electron transport chain* pathway, the five proteins predicted (UniProt ID: SMIM4_HUMAN, MOS1_HUMAN, A8MTT3_HUMAN, MANBL_HUMAN and SPAT9_HUMAN) were annotated as single-pass membrane proteins, and the 'transmembrane_domain' predicate was observed in 42 percent of the 77 proteins annotated in this pathway. Furthermore, the predicate 'protein_length_low' has appeared in all (100 percent) of the predicted proteins in the *Electron transport chain* pathway and in 78 percent of the annotated proteins in Reactome for this pathway; so, proteins in this pathway could be characterized by short protein sequences. Interestingly, SMIM4 and MOS1 are localized in mitochondrion according to UniProt annotations, where several enzyme complexes involved in the electron-transport system are anchored in place by transmembrane proteins. In addition, as a source of energy, mitochondria participate in other events like cell differentiation [48]. Although the mitochondrion annotation is not specified, another membrane-protein that was predicted, SPAT9, has "cellular differentiation" as a biological process annotation. By contrast, the A8MTT3 and MANBL are uncharacterized proteins.

Analysis of the *Telomere maintenance* pathway revealed a frequency of 100 percent for the predicate 'complex_interaction' and 'transcripts_low' in the predicted proteins (UniProt ID: HPGDS_HUMAN, CSN4_HUMAN, PIAS4_HUMAN, DTX1_HUMAN and APBP2_HUMAN). Moreover, these predicates were associated with 64 and 49 percent of the 45 annotated proteins in this pathway, respectively. This pathway was the only pathway predicted using the predicate 'transcripts_low', i.e., the predicted proteins were encoded by only one transcript. With the exception of HPGDS, a bi-functional enzyme (hematopoietic prostaglandin D synthase –EC 5.3.99.2–and glutathione S-transferase –EC 2.5.1.18–), the remaining four proteins all contained annotated nucleic acid binding motifs, such as: the Winged helix-turn-helix transcription repressor DNA-binding motif; the SAP-motif, a putative DNA binding motif found in diverse nuclear

proteins involved in chromosomal organization; the Zinc-finger domains now recognized to bind DNA, RNA, protein and/or lipid substrates; and the tetratrico peptide repeat region, which despite mediating protein-protein interactions and the assembly of multiprotein complexes in a wide-range of proteins, also adopts a helix-turn-helix arrangement commonly found in DNA-binding proteins (see InterPro annotations of these proteins).

"Telomeres are protein-DNA complexes at the ends of linear chromosomes that are important for genome stability" [7]. In humans, the mechanism of telomere replication remains poorly understood and further knowledge regarding transcriptional, translational and post-translational regulation of telomere-binding proteins is required [49].

"Ubiquitin (Ub) attachment principally regulates interactions with other macromolecules, such as proteasome-substrate binding or protein recruitment to chromatin" [50]. Furthermore, although several similarities are evident in pathways involved in activating and conjugating Ub and ubiquitin-like (Ubl) proteins to particular lysine residues within target proteins, the mechanism of exchange between small-Ubl-modifier (SUMO) proteins and the ubiquitination process remains unclear [50], [51].

Interestingly, UniProt annotations associated three of the RLE predicted proteins to Ub/Ubl conjugation pathway (CSN4, PIAS4 and DTX1), although Ub conjugation is involved in many eukaryotic cell processes. One of these three predicted proteins (CSN4) is a component of the COP9 signalosome complex and an essential regulator of the Ub-conjugation pathway in response to DNA damage [52]. PIAS4 is an E3 SUMO-protein ligase [53] and DTX1 also exhibits Ub-ligase activity *in vitro* [54]. Notably, the C-terminal domain (residues 946-1,132) of human telomerases is efficiently ubiquitinated *in vivo* by the E3-ligase MKRN1 and the ligase RING-finger domain is essential for the physical interaction between these proteins [55]. The predicted ligases PIAS4 and DTX1, as well as MKRN1, contain a RING-finger domain with conserved histidine and cysteine residues. Indeed, mutating His307Glu in the RING-finger domain of MKRN1 abolishes its ubiquitination activity [55]. Moreover, a connection between the maintenance of genome stability and the evolutionary conserved family of SUMO-targeted Ub-ligases has recently been proposed [56].

In the last example, we discuss the *Integrin cell surface interactions* pathway (partially represented in Fig. 6), where almost all of the five predicted proteins are cell surface receptors with a single-pass type I membrane architecture. The pathway shown includes some of the annotated proteins, their connections and their similarities with predicted proteins. The connector line indicates the proteins with the most similar predicates between RLE predictions and Reactome annotations (Fig. 6). According to this result we hypothesize that predicted proteins could have similar or related functions to the proteins annotated in Reactome: IL3RB_HUMAN is a cytokine receptor subunit B; CD22_HUMAN is the B-cell receptor CD22; NPHN_HUMAN is a specific cell adhesion receptor; and FPRP_HUMAN is a prostaglandin F2 receptor. CNTN1_HUMAN (contactin-1 or glycoprotein gp135) mediates cell surface interactions during nervous system development and is attached to the membrane by a lipid-anchor.

Taken together, these domain annotations and literature findings provide evidences that RLE *de novo* predictions have a biological sense.

3.6 Comparison with Pathway Prediction Based on Interaction Networks

We compared our RLE system with an alternative method used for pathway prediction that relies on molecular interaction network data [17].

This previous methodology, that expands pathways and other cellular processes, maps the proteins onto protein-protein interaction networks, expanding the pathway with densely interconnected interaction partners that increase pathway compactness. In this method the interaction network is the only input data used. Hence, the only candidates for expansion are proteins that interact directly with proteins pertaining to the original pathway and that fulfil a series of topological conditions [17].

Due to the novelty of the goal, RLE comparison to similar state-of-the-art methods is not easy. First, the ProtFun method [19], which uses sequence data alone, is a historical and unavailable method. This method uses a different approach, predicting GO terms instead of Reactome pathways as RLE does. A second method by Glaab et al. [17], which use interaction data alone, is not directly comparable to RLE in the same test performance terms, since it is not based on scores. Nevertheless, to illustrate the test performance of Glaab et al. method, we included a specific point in the PR and ROC curves (see Fig. 3), due to the lack of scores do not allow to represent a line. Fig. 3 shows how the Glaab et al. test performance point is under the RLE test performance line. For a proper comparison between RLE and Glaab et al. method, we evaluated these two methods with the application data set, described in Section 2.3.2. Glaab et al. method needs our complete pathways (instead of the network of non-redundant proteins in sequence-similarity terms) as input in order to suitably compute its topological metrics. Thus, Glaab et al. method predicted proteins in 29 of the 37 pathways with a total of 150 directly connected proteins, 90 different proteins (60 percent of the 150 proteins that were added overall). Therefore, the prediction coverage is greater in our RLE approach (383 proteins) than in this earlier method (150 proteins).

These two methods associated new proteins to 21 common pathways, although for each individual pathway there were very few common proteins that were predicted by both methods. Specifically, we detected only five common proteins using the two methods: two in the *Gene Expression* (PathwayID:35) and three in the *Transcription* (PathwayID:21). Furthermore, if we consider proteins predicted by the two methods in different pathways, neither produced an increase in the number of coincidences (for overlaps see Supplementary Fig. S1, available online). Accordingly, TAF2_HUMAN, RPC3_HUMAN and B3KRR0_HUMAN were proposed to be related to the *Transcription* pathway, while MED29_HUMAN and PDCD4_HUMAN were related to the *Gene Expression* pathway. TAF2_HUMAN is the transcription initiation factor TFIID subunit 2, while RPC3_HUMAN is the DNA-directed RNA polymerase III subunit C3, both proteins that are located in the nucleus. B3KRR0_HUMAN is an

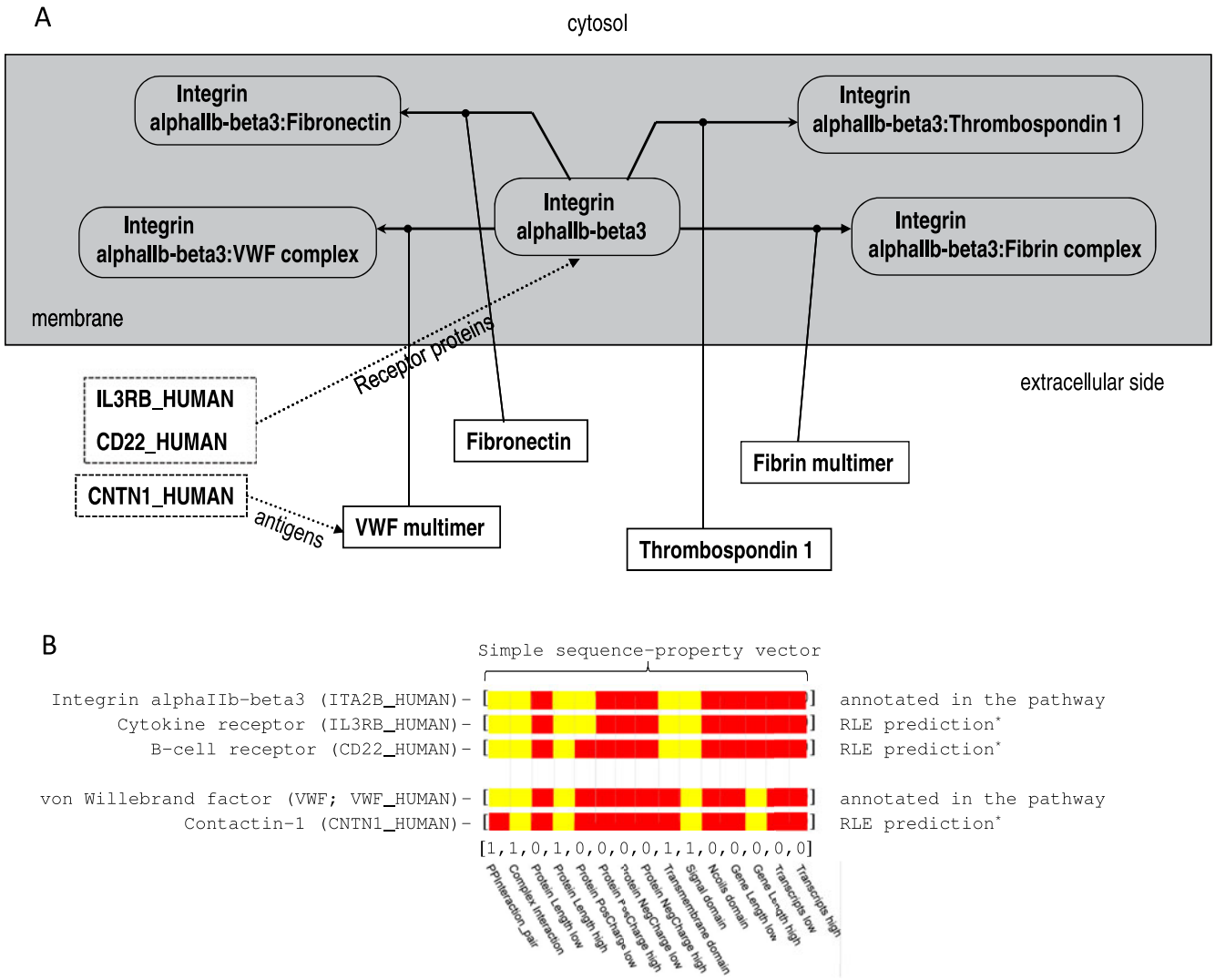

Fig. 6. A hypothetical diagram of the human Integrin cell surface interactions pathway defined by the RLE system. In the diagram (panel A), some of the annotated proteins in the pathway, their connections and three RLE predicted proteins are represented. The dashed lines represent proteins predicted by the RLE system. Panel B shows a comparison of the simple sequence-property vectors for the annotated and predicted proteins: yellow represents true (1) and red represents false (0). The numerical vector shown is the mode of the above five coloured vectors above. *Cytokine receptor common subunit beta (IL3RB_HUMAN) is a high affinity receptor for interleukin-3, interleukin-5 and granulocyte-macrophage colony-stimulating factor; the B-cell receptor CD22 (CD22_HUMAN) mediates interactions between B-cells; contactin-1 (CNTN1_HUMAN) mediates cell surface interactions during nervous system development.

uncharacterized protein that has strong similarity to the DNA excision repair protein ERCC-1. In the second pathway, MED29_HUMAN is a mediator of RNA polymerase II transcription subunit 29, and PDCD4_HUMAN is the programmed cell death protein 4 that inhibits translation initiation by binding to eukaryotic initiation factor 4A (eIF4A), as well as inhibiting the helicase activity of eIF4A. While the biological relevance of these findings requires further study, UniProt annotation of these proteins and their simultaneous prediction by two independent methods could increase the reliability of these results.

In addition, we used a functional semantic similarity measure [57], [58] in order to compare both pathway prediction methods on the basis of GO Biological Process annotations. We use Jiang and Conrath's similarity measure [57], with GO Biological Process terms (from Ensembl Release 56, all evidence codes except ISS). We computed the best-match average similarity measure [58] between all pair-wise combinations of proteins, obtaining a semantic similarity value for each pathway. This semantic similarity between the original pathway proteins and the predicted proteins was stronger for the method of Glaab et al. than for our RLE approach. The average similarity according to the number of predicted pathways in each system was 0.700 using the method of Glaab et al. for 29 pathways and 0.591 with the RLE for 28 pathways. Moreover, the proteins predicted by each approach were more semantically similar to the original pathway than to each other (0.412 for Glaab et al. versus RLE), demonstrating that the proteins identified by both methods differ significantly.

The overlap between proteins related to different pathways using the RLE approach was lower than that found using the molecular interaction network based method [17], as 15 percent of proteins were associated to more than one pathway in the RLE system as opposed to approximately 30 percent using the latter method (see overlaps in

Supplementary Fig. S1, available online). If we ignore this overlap (i.e., proteins predicted to be related to more than one pathway), both methods are closer to the original pathway in semantic similarity terms. Taken together, these findings suggest that the RLE approach provides better results in 15 pathways, while the method using only molecular interaction networks is superior in another 15 pathways. Without taking into account the overlap, our RLE system associated proteins to only 1 pathway less (27 different pathways), while the Glaab et al. method stopped predicting some proteins in 10 pathways (that is 19 different predicted pathways). Thus, this result confirms that the prediction method presented by Glaab et al. is limited to a small functional area (i.e., proteins highly connected in specific pathways), while the proteins predicted by RLE highlight more distant relationships and they are related to a wider functional realm.

We conclude that the two pathway prediction methods are complementary given the different number of proteins predicted by the two systems, the small overlap between the proteins predicted by both methods, and the distinct distances in terms of the exploration space. This means that while the expansion with the Glaab et al. method involves few proteins connecting many pathways, RLE predicts many proteins that are different for each pathway.

3.7 Application to New Releases

We have applied the RLE system to the new releases of the Reactome and the rest of the input databases. In the case of Reactome release 40, Ensembl release 67 and BioGrid release 3.1.89, RLE finds proteins with similar characteristics to the original pathway in 32 of the previously analyzed 37 pathways. According to the molecular diversity per pathway, RLE applies several rules to predict proteins related to the Reactome pathway proteins in 24 of these pathways. RLE predicts 572 proteins associated to pathways, with 91 of them predicted to be associated to more than one pathway, and thus, only 16 percent of the predicted proteins connect two or more pathways. The list of specific proteins with similar properties to each pathway is available on the RLE web (http://rle.bioinfo.cnio.es), where the prediction of the human proteome will appear for future releases of the Reactome pathway database.

4 Discussion and Conclusions

This study describes a system that predicts pathway associations based on a function prediction approach that relies on combinations of simple properties associated with each protein. The predictions are based mainly on sequence features (including the number of isoforms), independent of the existence of homologies, which means this method can be applied to poorly characterized proteins. The predictions also consider some properties related to the position of the proteins in protein-protein interaction networks and protein complexes (i.e., interaction partners and their corresponding features). This relational information distinguishes this system from others based only on individual characteristics. Using this approach, we searched for specific proteins with molecular similarities to pathway fragments rather than proteins with characteristics common to the overall pathway at the biological process level.

Since it is an approach that searches for proteins with similar simple properties to pathway proteins, its novelty involves the difficulty of finding an established framework as a reference, such as secondary structure prediction or protein-protein interaction prediction. In turn, this fact makes it difficult to prove the results obtained with additional data and to compare this approach with other applications.

We are aware that comprehensive validation is always a key step in the evaluation of machine learning methods. However, the validation strategy has to be adequate to the characteristics of the problem and data sets. For example, as described in Section 2.3.1, a typical 10-CV is not suitable to the small and diverse data set typical of the problem described here, since proteins assigned to Reactome pathways can not be split in consistent, equivalent and unbiased partitions. In other words a multi-class and multi-label problem with many small data sets is not appropriate for this 10-CV strategy. On the other hand, a comparison with BLAST results it is not appropriated, since RLE and BLAST are based on different assumptions. RLE includes additional input data different from amino-acid sequences (i.e., number of transcript, gene length, protein-protein interactions and complexes) that BLAST doesn't take into account and, consequently, the predictions wouldn't be comparable. Therefore, the adequate validation strategy in this case is the partition of the data sets in large (2:1) sets, and the systematic comparison with the most similar available method (Glaab et al. method). We have also carried out a detailed interpretation of the biological relevance of the results of the prediction method.

By using a relational representation and applying the new RLE system, we have associated 383 uncharacterized proteins to 28 human Reactome pathways, given the specific chosen cut-offs (see Section 2.4). As each RLE prediction has an associated score, a more restricted cut-off could be chosen, taking into account that several predicted proteins share the same score (i.e., they are classified by the same decision tree branch). The level of predicted proteins differs from pathway to pathway in terms of performance and of the different molecular properties of the proposed associations. RLE enhances the annotations of both groups of proteins, those predicted to be related to the pathway and those originally annotated in the pathway.

Regarding the use of sequence and interaction features, the prediction results indicate that interactions are not the only useful feature in the learning process, as interactions alone do not reach performance values as high as the complete system. Therefore, the results we obtained support the use of sequence features in addition to interaction information.

As expected, our proposed proteins associated to pathways differed from those predicted by the previous method based only on interaction networks [17], both these methods being complementary. The RLE system proposed here predicts proteins with more diverse functions, searching within and beyond the proximal space (for example, less proteins in the intersection of pathways). Indeed, as this RLE system focuses on pathway-specific proteins rather than on the connections between pathway proteins, the prediction avoids overlap between different pathways.

Proteins predicted by RLE provide alternative hypothesis for some of the cellular processes studied. Of particular

note were the proteins predicted to be related to *Transcription*, *Gene expression*, *Electron transport chain*, *Telomere maintenance* and *Integrin cell surface interactions* pathways. In addition, combining UniProt annotations and literature findings with RLE results slightly augments the reliability of the relationship of a predicted protein to a specific pathway. These results also confirm that in terms of the biological implications of specific proteins, a low AUPRC value (such as that obtained for the *Telomere maintenance* pathway) does not always indicate a poor prediction.

Finally, the sophisticated Relational Learning-based Annotation procedure may be applied to predict proteins with similar properties to some other pathway databases when the annotated proteins available fulfil the learning requirements. Furthermore, this system may be employed for the functional annotation of unknown genes and proteins with a different vocabulary, even permitting data to be shared that have already been represented relationally in our knowledgebase.

ACKNOWLEDGMENTS

The authors wish to thank Anaïs Baudot and Enrico Glaab for their interesting discussions, the ACE tool developers (especially Daan Fierens) for their valuable help with the learning tool, and the anonymous referees and editors for their valuable suggestions to improve the manuscript. This work was supported by the IMI Innovative Medicines Initiative under the Open PHACTS project (grant agreement 115191), by the Spanish Ministry of Science and Innovation under the Supercomputation and eScience (SyeC) CONSOLIDER project (grant agreement CSD2007-00050) and the BIO2012-40205 and TRA2011-29454-C03-03 projects, and by the Ambient Assisted Living Programme under the Trainutri project (AAL-2009-2-129).

REFERENCES

[1] M. P. Cary, G. D. Bader, and C. Sander, "Pathway information for systems biology," *FEBS Lett.*, vol. 579, no. 8, pp. 1815–1820, 2005.

[2] H. S. Ooi, G. Schneider, T.-T. Lim, Y.-L. Chan, B. Eisenhaber, and F. Eisenhaber, *Biomolecular Pathway Databases*, New York, NY, USA, Humana Press, vol. 609, pp. 129–144, 2010.

[3] M. K. Sakharkar, K. R. Sakharkar, and S. Pervaiz, "Druggability of human disease genes," *Int. J. Biochem. Cell B.*, vol. 39, no. 6, pp. 1156–1164, 2007.

[4] E. Demir, M. P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P.~D Eustachio, C. Schaefer, J. Luciano, F. Schacherer, I. Martinez-Flores, Z. Hu, V. Jimenez-Jacinto, G. Joshi-Tope, K. Kandasamy, A. Lopez-Fuentes, H. Mi, E. Pichler, I. Rodchenkov, A. Splendiani, S. Tkachev, J. Zucker, G. Gopinath, H. Rajasimha, R. Ramakrishnan, I. Shah, M. Syed, N. Anwar, O. Babur, M. Blinov, E. Brauner, D. Corwin, S. Donaldson, F. Gibbons, R. Goldberg, P. Hornbeck, A. Luna, P. Murray-Rust, E. Neumann, O. Reubenacker, M. Samwald, M. van Iersel, S. Wimalaratne, K. Allen, B. Braun, M. Whirl-Carrillo, K.-H. Cheung, K. Dahlquist, A. Finney, M. Gillespie, E. Glass, L. Gong, R. Haw, M. Honig, O. Hubaut, D. Kane, S. Krupa, M. Kutmon, J. Leonard, D. Marks, D. Merberg, V. Petri, A. Pico, D. Ravenscroft, L. Ren, N. Shah, M. Sunshine, R. Tang, R. Whaley, S. Letovksy, K. H. Buetow, A. Rzhetsky, V. Schachter, B. S. Sobral, U. Dogrusoz, S. McWeeney, M. Aladjem, E. Birney, J. Collado-Vides, S. Goto, M. Hucka, N. L. Novere, N. Maltsev, A. Pandey, P. Thomas, E. Wingender, P. D. Karp, C. Sander, and G. D. Bader, "The BioPAX community standard for pathway data sharing," *Nat. Biotech.*, vol. 28, no. 9, pp. 935–942, 2010.

[5] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander, "Pathway Commons, a web resource for biological pathway data," *Nucleic Acids Res.*, vol. 39, no. suppl 1, pp. 685–690, 2011.

[6] G. D. Bader, M. P. Cary, and C. Sander, "Pathguide: a pathway resource list," *Nucleic Acids Res.*, vol. 34, no. suppl 1, pp. 504–506, 2006.

[7] L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio, "Reactome knowledgebase of human biological pathways and processes," *Nucleic Acids Res.*, vol. 37, no. suppl 1, pp. 619–622, 2009.

[8] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic Acids Res.*, vol. 40, no. D1, pp. 109–114, 2012.

[9] R. Caspi, T. Altman, J. M. Dale, K. Dreher, C. A. Fulcher, F. Gilham, P. Kaipa, A. S. Karthikeyan, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, S. Paley, L. Popescu, A. Pujar, A. G. Shearer, P. Zhang, and P. D. Karp, "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases," *Nucleic Acids Res.*, vol. 38, no. suppl 1, pp. 473–479, 2010.

[10] T. Kelder, M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo, and A. R. Pico. (2012) Wikipathways: Building research communities on biological pathways. *Nucleic Acids Res.* [Online]. *40(D1)*, pp. D1301–D1307. Available: http://nar.oxfordjournals.org/content/40/D1/D1301.abstract

[11] L. J. Lu, A. Sboner, Y. J. Huang, H. X. Lu, T. A. Gianoulis, K. Y. Yip, P. M. Kim, G. T. Montelione, and M. B. Gerstein, "Comparing classical pathways and modern networks: towards the development of an edge ontology," *Trends Biochem. Sci.*, vol. 32, no. 7, pp. 320–331, 2007.

[12] T. Korcsmaros, M. S. Szalay, P. Rovo, R. Palotai, D. Fazekas, K. Lenti, I. J. Farkas, P. Csermely, and T. Vellai, "Signalogs: Orthology-based identification of novel signaling pathway components in three metazoans," *PLoS One*, vol. 6, no. 5, p. 19240, 2011.

[13] H. Frohlich, M. Fellmann, H. Sültmann, A. Poustka, and T. Beissbarth, "Predicting pathway membership via domain signatures," *Bioinformatics*, vol. 24, no. 19, pp. 2137–2142, 2008.

[14] P. D. Karp, S. Paley, and P. Romero, "The pathway tools software," *Bioinformatics*, vol. 18, no. suppl 1, pp. 225–232, 2002.

[15] M. E. Adriaens, M. Jaillard, A. Waagmeester, S. L. M. Coort, A. R. Pico, and C. T. A. Evelo, "The public road to high-quality curated biological pathways," *Drug Disc. Today*, vol. 13, no. 19–20, pp. 856–862, 2008.

[16] K. L .J. Prather and C. H. Martin, "De novo biosynthetic pathways: Rational design of microbial chemical factories," *Current Opinion Biotechnol.*, vol. 19, no. 5, pp. 468–474, 2008.

[17] E. Glaab et al., "Extending pathways and processes using molecular interaction networks to analyse cancer genome data," *BMC Bioinf.*, vol. 11, no. 1, article 597, 2010.

[18] L. J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, C. Kesmir, H. Nielsen, H. H. Staerfeldt, K. Rapacki, C. Workman, C. A. Andersen, S. Knudsen, A. Krogh, A. Valencia, and S. Brunak, "Prediction of human protein function from post-translational modifications and localization features," *J. Mol. Biol.*, vol. 319, no. 5, pp. 1257–1265, 2002.

[19] L. J. Jensen, R. Gupta, H. H. Staerfeldt, and S. Brunak, "Prediction of human protein function according to Gene Ontology categories," *Bioinformatics*, vol. 19, no. 5, pp. 635–642, 2003.

[20] J. D. Bendtsen, L. J. Jensen, N. Blom, G. von Heijne, and S. Brunak, "Feature-based prediction of non-classical and leaderless protein secretion," *Protein Eng. Design Selection*, vol. 17, no. 4, pp. 349–356, 2004.

[21] A. Clare, A. Karwath, H. Ougham, and R. D. King, "Functional bioinformatics for Arabidopsis thaliana," *Bioinformatics*, vol. 22, no. 9, pp. 1130–1136, 2006.

[22] C. Vens, J. Struyf, L. Schietgat, S. Dzeroski, and H. Blockeel, "Decision trees for hierarchical multi-label classification," *Mach. Learn.*, vol. 73, no. 2, pp. 185–214, 2008.

[23] G. Wu, X. Feng, and L. Stein, "A human functional protein interaction network and its application to cancer data analysis," *Genome Biol.*, vol. 11, no. 5, p. 53, 2010.

[24] B. Jassal, "Pathway annotation and analysis with Reactome: The solute carrier class of membrane transporters," *Human Genomics*, vol. 5, no. 4, pp. 310–315, 2011.
[25] S. Dzeroski and N. Lavrac, *Relational Data Mining*. New York, NY, USA: Springer, 2001.
[26] J. E. Gewehr, M. Szugat, and R. Zimmer, "BioWeka extending the Weka framework for bioinformatics," *Bioinformatics*, vol. 23, no. 5, pp. 651–653, 2007.
[27] T. J. P. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle, and P. Flicek, "Ensembl 2009," *Nucleic Acids Res.*, vol. 37, no. suppl 1, pp. 690–697, 2009.
[28] D. Smedley, S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, and A. Kasprzyk, "BioMart - Biological queries made easy," *BMC Genomics*, vol. 10, article 22, 2009.
[29] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: A general repository for interaction datasets," *Nucleic Acids Res.*, vol. 34, no. suppl 1, pp. 535–539, 2006.
[30] A. Chatr-Aryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli, and G. Cesareni, "MINT: The Molecular INTeraction database," *Nucleic Acids Res.*, vol. 35, no. suppl 1, pp. 572–574, 2007.
[31] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler, "IntAct: An open source molecular interaction database," *Nucleic Acids Res.*, vol. 32, no. database issue, pp. 452–455, 2004.
[32] S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. K. B. Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, H. N. Shivashankar, B. P. Rashmi, M. A. Ramya, Z. Zhao, K. N. Chandrika, N. Padma, H. C. Harsha, A. J. Yatish, M. P. Kavitha, M. Menezes, D. R. Choudhury, S. Suresh, N. Ghosh, R. Saravana, S. Chandran, S. Krishna, M. Joy, S. K. Anand, V. Madavan, A. Joseph, G. W. Wong, W. P. Schiemann, S. N. Constantinescu, L. Huang, R. Khosravi-Far, H. Steen, M. Tewari, S. Ghaffari, G. C. Blobe, C. V. Dang, J. G. N. Garcia, J. Pevsner, O. N. Jensen, P. Roepstorff, K. S. Deshpande, A. M. Chinnaiyan, A. Hamosh, A. Chakravarti, and A. Pandey, "Development of human protein reference database as an initial platform for approaching systems biology in humans," *Genome Res.*, vol. 13, no. 10, pp. 2363–2371, 2003.
[33] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, 1990.
[34] U. Hobohm, M. Scharf, R. Schneider, and C. Sander, "Selection of representative protein data sets," *Protein Sci.*, vol. 1, no. 3, pp. 409–417, 1992.
[35] S. Griep and U. Hobohm, "PDBselect 1992-2009 and PDBfilterselect," *Nucleic Acids Res.*, vol. 38, no. suppl 1, pp. 318–319, 2010.
[36] O. Emanuelsson, H. Nielsen, and G. V. Heijne, "ChloroP, A neural network-based method for predicting chloroplast transit peptides and their cleavage sites," *Protein Sci.*, vol. 8, no. 5, pp. 978–984, 1999.
[37] K. Wang, D. W. Ussery, and S. Brunak, "Analysis and prediction of gene splice sites in four Aspergillus genomes," *Fungal Genetics Biol.*, vol. 46, no. suppl 1, pp. 14–18, 2009.
[38] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, vol. 147, no. 1, pp. 195–197, 1981.
[39] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.
[40] R. Kohavi. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection *Proc. 14th Int. Joint Conf. Artif. Intell. - vol. 2*, pp. 1137–1143. [Online]. Available: http://dl.acm.org/citation.cfm?id=1643031.1643047
[41] L. Dehaspe and L. D. Raedt, "Mining association rules in multiple relations," in *Proc. 7th Int. Workshop Inductive Logic Programm.*, 1997, pp. 125–132.
[42] H. Blockeel, L. Dehaspe, J. Ramon, J. Struyf, A. V. Assche, C. Vens, and D. Fierens. (2006). The ACE Data Mining System. User's Manual. [Online]. Available: http://www.cs.kuleuven.be/~dtai/ACE.
[43] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast discovery of association rules," in *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA, USA: MIT Press, 1996, pp. 307–328.
[44] H. Blockeel, L. D. Raedt, and J. Ramon, "Top-down induction of clustering trees," in *Proc. 15th Int. Conf. Mach. Learn.*, 1998, pp. 55–63.
[45] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.
[46] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 233–240.
[47] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, 2002.
[48] H. M. McBride, M. Neuspiel, and S. Wasiak, "Mitochondria: More than just a powerhouse," *Current Biol.*, vol. 16, no. 14, pp. 551–560, 2006.
[49] Y. Xu, "Chemistry in human telomere biology: Structure, function and targeting of telomere DNA/RNA," *Chem. Soc. Rev.*, vol. 40, no. 5, pp. 2719–2740, 2011.
[50] M. Hochstrasser, "Origin and function of Ubiquitin-like proteins," *Nature*, vol. 458, no. 7237, pp. 422–429, 2009.
[51] D. Bailey and P. O'Hare, "Comparison of the SUMO1 and Ubiquitin conjugation pathways during the inhibition of proteasome activity with evidence of SUMO1 recycling," *Biochem. J.*, vol. 392, no. 2, pp. 271–281, 2005.
[52] R. Groisman, J. Polanowska, I. Kuraoka, J.-I. Sawada, M. Saijo, R. Drapkin, A. F. Kisselev, K. Tanaka, and Y. Nakatani, "The Ubiquitin ligase activity in the DDB2 and CSA complexes is differentially regulated by the COP9 signalosome in response to DNA damage," *Cell*, vol. 113, no. 3, pp. 357–367, 2003.
[53] M. Ihara, H. Yamamoto, and A. Kikuchi, "SUMO-1 modification of PIASy, an E3 ligase, is necessary for PIASy-dependent activation of Tcf-4," *Mol. Cell. Biol.*, vol. 25, no. 9, pp. 3506–3518, 2005.
[54] K. Takeyama, R. C. T. Aguiar, L. Gu, C. He, G. J. Freeman, J. L. Kutok, J. C. Aster, and M. A. Shipp, "The BAL-binding protein BBAP and related Deltex family members exhibit Ubiquitin-protein Isopeptide ligase activity," *J. Biol. Chem.*, vol. 278, no. 24, pp. 21 930–21 937, 2003.
[55] J. H. Kim, S. M. Park, M. R. Kang, S. Y. Oh, T. H. Lee, M. T. Muller, and I. K. Chung, "Ubiquitin ligase MKRN1 modulates telomere length homeostasis through a proteolysis of hTERT," *Genes Dev.*, vol. 19, no. 7, pp. 776–781, 2005.
[56] S. Nagai, N. Davoodi, and S. M. Gasser, "Nuclear organization in genome stability: SUMO connections," *Cell Res.*, vol. 21, no. 3, pp. 474–485, 2011.
[57] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proc. Int. Conf. Res. Comput. Linguistics*, 1997, pp. 19–33.
[58] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, and F. M. Couto, "Semantic similarity in biomedical ontologies," *PLoS Comput. Biol.*, vol. 5, no. 7, p. e1000443, 2009.

Beatriz García-Jiménez received the PhD degree in computer science (Bioinformatics) from the Universidad Carlos III de Madrid (UC3M) in 2012, under the supervision of Araceli Sanchis (UC3M) and Alfonso Valencia (CNIO). She was a assistant professor of the Computer Science Department at UC3M, from 2006 to 2013. From February 2013 to June 2013, she was a visiting postdoctoral researcher of the computational biology and applied algorithmics department at Max-Planck Institute for Informatics (MPII). Since September 2013, she has been a postdoctoral fellow in Biological Informatics group at Centre for Plant Biotechnology and Genomics UPM-INIA. Her thesis won the 2012 National Best Thesis Award, in experimental and technological sciences area, from the Royal Spanish Doctoral Academy. Her research interests include functional annotation, knowledge relational representation, machine learning and protein networks.

Tirso Pons received the PhD degree in biology from the University of Havana in 2003. From 2003-2005, he was a postdoctoral researcher at the Center for Genetic Engineering and Biotechnology (CIGB) and from 2006-2011, an associate professor in Bioinformatics at Faculty of Biology, University of Havana. He was a visiting scientist at Protein Design group (CNB-CSIC), Autonomous University of Madrid, and at Biocomputing Unit, European Molecular Biology Laboratory (EMBL) in Heidelberg. He is a member of the editorial board of *Bioinformatics and Biology Insights*, and *Journal of Proteome Science & Computational Biology*. Since 2011, he has been a staff scientist at the Structural Biology and Biocomputing Programme of the Spanish National Cancer Research Centre (CNIO), Madrid. His research interests include sequence analysis, protein interactions, and protein three-dimensional structure and functional residues prediction.

Araceli Sanchis received the PhD degree in computer science from UPM in 1990. She received the BS degree in chemistry from the Complutense University of Madrid in 1991 and the other PhD degree in physical chemistry from Complutense University of Madrid in 1994. She is a university associate professor of computer science at Universidad Carlos III de Madrid (UC3M) since 1999. She has been vicedean of the Computer Science degree at UC3M and, currently, she is head of the CAOS group (Grupo de Control, Aprendizaje y Optimizacin de Sistemas), based on machine learning and optimization. She has published more than 130 journal and conference papers mainly in the field of machine learning.

Alfonso Valencia is a vicedirector of basic research and director of the Structural Biology and Biocomputing Programme at the Spanish National Cancer Research Centre (CNIO). He is also a director of Spanish Bioinformatics Institute (INB-ISCIII). Biologist by training with PhD degree in molecular biology, his research is based on the use of computation for the analysis of large collection of genomic information with particular emphasis in the study of protein families and protein interaction networks. In this context, he has dedicated considerable efforts to integrate method development in different area of bioinformatics, from comparative genomics to text mining. His recent work focuses in the area of cancer genomics. He is a president of the International Society for Computational Biology (2015-2018). He is elected member of the European Molecular Biology Organization and Professor Honoris Causa of the Danish Technical University DTU. As an executive editor of *Bioinformatics*, published by OUP he has promoted the integration of computational technologies in areas of molecular biology and biomedicine.