# Integrated Prediction of Protein Function, Interactions and Pathways with Statistical Relational Learning

**Principal Investigator**
**Andrea Passerini** (passerini@disi.unitn.it), University of Trento, Italy

**Google sponsors**
**Massimiliano Ciaramita** (massi@google.com)
**Diego Federici** (diegofederici@google.com)

**Abstract**
Accurate determination of protein function, interactions, and biological pathways has extensive applications in biology, with a direct impact on the effectiveness of drug/enzyme design. The creation of a complete database of protein features would significantly advance our understanding of biological processes, shedding light on disease development and aging (the main goal of Google's new enterprise Calico[1]). Unfortunately, the creation of such database would take decades, because of the extremely expensive and time consuming experimental methods involved. A more scalable (and extensively applied) approach is to use machine learning techniques to predict these features [1]. However, prediction accuracy is an issue with current methodologies, for two main limitations. First, the intrinsic relational nature of the problem is often neglected. For instance, protein–protein interactions can be seen as a hierarchical process occurring at three related levels: *proteins* bind by means of specific *domains*, which in turn form interfaces through patches of *residues*. A collective and consistent prediction of interactions at all different levels is crucial for achieving reliable results. Second, machine learning approaches have been typically applied on a per-database, *ad hoc* manner, thus neglecting the correlation between information stored in distinct databases. Combining these sources of information is a challenging task, as biological databases — such as UniProt [2], BioGRID [3], KEGG [4] — are semi-manually compiled by domain experts according to often incompatible schemas, based on discordant definitions of the same biological concept (such as gene function), making interoperability and integration difficult. In an effort to lessen this issue, recently researchers have developed and applied a number of ontologies to formally define the database semantics, culminating with the OBO initiative [5]. These developments lower the barrier for performing automatic reasoning (i.e. *automatic compilation* of entries and *consistency enforcement*) over distinct, yet correlated, resources. We propose to develop a method to *seamlessly predict the missing entries in multiple databases*. Our proposal is based on Semantic Based Regularization [6], a state-of-the-art *statistical relational learning* method [7], that performs collective classification over attribute-value representations using (weighted) first-order logic rules. The latter encode the relations defined by the ontologies, and allow to inject domain knowledge into the learning task. We plan to apply the method to the prediction of protein interactions, protein function and biological pathways, in an aggregate manner. We also provide preliminary results on one of these tasks, which should help to evaluate the feasibility of our proposal.

**Goals** To develop a method for the collective prediction of multiple inter-related protein properties with ontology- and knowledge-driven rules.

**Detailed plan**
The focus of this proposal is on the collective prediction of multiple interrelated protein properties stored in distinct biological databases. More specifically, our prediction targets are protein function, protein–protein interactions, and signal transduction pathways. Protein *function* determines the biological role of a protein, e.g. whether it is an enzyme, an antibody, a transport protein, *etc.* Functions are structured as a hierarchy of controlled terms, as defined by the Gene Ontology (GO) [8]. The

[1] Calico official announcement, http://googlepress.blogspot.gr/2013/09/calico-announcement.html, September, 2013.

GO also encompasses sub-cellular localization information, i.e. where a protein is supposed to reside within the cell, e.g. in the nucleus, cytosol, mitochondrion, *etc.* In order to carry out their function, most proteins *interact* with (bind to) other proteins; a group of proteins and their interactions is called a protein–protein interaction network (PIN). PINs are intimately relational, and can be viewed as graphs, with proteins as nodes and interactions as edges. *Pathways* are evolutionarily conserved sub-modules of the PIN that are specialized for a particular function. In particular, signal transduction pathways consist of groups of interacting proteins that collectively transport information about events in the cell, e.g. apoptosis (programmed cell death).

These three fundamental aspects of cell life are stored in distinct, independent biological databases. Notwithstanding the prolonged effort of domain experts (who carry out the biological experiments, compile the annotations, and validate the results), the databases are far from complete. Machine learning methods have been extensively applied to the automatic completion of the missing entries, by leveraging the large amount of existing data. This line of research produced methods to predict several protein properties — secondary and tertiary structure, solvent accessibility, active sites, disorder, function, interactions, *etc.* [1] — from the raw amino acid sequence and additional *known* information. Said properties are heavily correlated, e.g. interacting proteins carry out the same biological function, and proteins that reside in different parts of the cell can not interact. Current methods however tend to predict only one property at a time, leading to inconsistent, sub-optimal predictions. The ability to integrate known relational information and perform collective predictions over the whole set of correlated properties could allow to substantially improve the accuracy of the annotations.

Semantic Based Regularization (SBR) [6] is a method able to solve this kind of task. It is a state-of-the-art *statistical relational learning* method [7] that handles low-level attribute-value representations of the objects (e.g. proteins and protein pairs) by leveraging the classical kernel machine framework, and allows to explicitly model relations between the objects and their properties using (soft) First Order Logic (FOL) constraints. Each target property is encoded as a kernel function, and the resulting multi-task objective function is conditioned by the FOL rules. In order to assess the abilities of SBR, we applied it to the prediction of protein–protein interactions [9]. PPIs occur at three related levels: *proteins* bind by means of specific *domains* (conserved motifs that occur frequently in protein sequences), which in turn form interfaces through patches of *residues* (amino acids). Consistency between different levels is based on two observations: if two proteins interact, at least two of their domains must interact, which in turn implies that some of their residues interact — and vice versa. We encoded these constraints as FOL rules, and used SBR to compute the binding state for all protein, domain and residue pairs, collectively. The results are very encouraging, as SBR largely outperformed the competition in *all* experimental settings[2]. Our plan is to generalize this approach to jointly predict functions, interactions and biological pathways, following the lessons learned in a similar multi-prediction integration task [10]. In the remainder, we outline our plan in three incremental steps.

A prerequisite step involves the collection of a representative dataset, sufficiently large to successfully train the learning method and accurately validate its results. In particular, we will extract the protein interactions from BioGRID, a large repository of experimentally confirmed and computationally predicted PINs from various organisms, assigning a different confidence to interactions depending on their source. We will also aggregate additional information about domain-level and residue-level interactions. Protein function and sub-cellular localization will be taken from UniPROT, an annotated catalog of all known protein sequences. Signal transduction pathways will be taken from KEGG, the Kyoto Encyclopedia of Genes and Genomes, which relies on its own ontology. The resulting dataset will be augmented with low-level representations of the proteins (including representative features such as sequence, conservation, and homology) taken from UniPROT and other resources. In order for the dataset to be as authoritative a benchmark as possible, we will initially focus on *S. Cerevisiae* (Baker's yeast), the best annotated model organism, and later move to *H. Sapiens*. In the spirit of open research, the resulting dataset will be made freely available.

As a second step, we plan to augment the current experiment with protein function prediction, i.e. prediction of GO terms, including subcellular localization information. As explained above, these properties are hierarchical, i.e. they form a *rooted tree* representing a general-to-specific relation. Predictions therefore consist of multiple, interdependent labels, one for each node in the tree. The main constraint is that, if a node $u$ is predicted true, so must be all nodes on the path from $u$ to the

[2] Details available at `https://sites.google.com/site/semanticbasedregularization/home/software/proteins_interaction`.

root. We plan to integrate the two learning tasks by introducing FOL constraints between function and interactions, according to these observations: i) interacting proteins often share the same function; ii) proteins residing in different sub-cellular locations are not likely to interact. Moreover, there is a strong correlation between localization and function (e.g. proteins that manipulate the DNA are likely to reside in the nucleus), and function and interactions (e.g. certain enzymes lose their catalytic ability when bound). One major challenge is that some proteins may exhibit multiple functions, and may reside in multiple cellular components. We plan to explicitly take this ambiguity into account by using soft (non deterministic) constraints. The resulting method would be able to *collectively* predict function, localization and interactions for a whole set of proteins — a feature never achieved before — and would guarantee the outputs to be consistent as a whole.

In a third step, we plan to extend the above experiment to include signal transduction pathway information. Biological pathways are typically mined from experimentally validated, functionally annotated PINs, in a semi-manual fashion. Pathways can be exploited in two manners. On one hand, since most organisms and cell types rely on the same basic biological functions, they also share the same fundamental pathways. Given a database of known pathways, we can constrain the predicted interaction/function network to include/exclude said pathways. This is indeed not an easy task, as it requires to develop a way to perform (soft) labeled sub-graph matching. The problem can be reduced by extracting (relatively small) common motifs from the pathways, which capture the most frequent function-interaction patterns, and encoding them as weighted FOL rules. On the other hand, by jointly predicting both the annotations (function, interactions) and the pathways, SBR can condition the three components in order to maximize the overall agreement.

The proposed system would be able to collectively infer the state of individual and network properties of proteins (function, interactions, and signaling pathways) for a whole organism. Thanks to the user-provided FOL rules, the predictions would be naturally consistent with each other and with the domain knowledge. Such a system would be an excellent tool to automatically enrich the current biological databases. Moreover, the system is highly modular, allowing to seamlessly integrate additional prediction tasks (i.e. protein properties) with only moderate effort.

**Data Policy:** the constructed database of protein features and relations will be released to the research community with no restrictions.

**Budget details**: financial support for one student for one year: 45,000 USD, travel and conferences: 3,000 USD. Total grand requested: 48,000 USD.

# References


[1] Pedro Larrañaga, Borja Calvo, Roberto Santana, et al. Machine learning in bioinformatics. *Briefings in bioinformatics*, 7(1):86–112, 2006.

[2] Amos Bairoch, Rolf Apweiler, Cathy H Wu, et al. The universal protein resource (uniprot). *Nucleic acids research*, 33(suppl 1):D154–D159, 2005.

[3] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, et al. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539, 2006.

[4] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[5] Barry Smith, Michael Ashburner, Cornelius Rosse, et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.

[6] Michelangelo Diligenti, Marco Gori, Marco Maggini, and Leonardo Rigutini. Bridging logic and kernel machines. *Machine learning*, 86(1):57–88, 2012.

[7] Lise Getoor and Ben Taskar. *Introduction to Statistical Relational Learning*. 2007.

[8] Michael Ashburner, Catherine A Ball, Judith A Blake, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[9] Claudio Saccà, Stefano Teso, Michelangelo Diligenti, and Andrea Passerini. Improved multi-level protein–protein interaction prediction with semantic-based regularization. *BMC Bioinformatics*, [Submitted].

[10] Stefano Teso and Andrea Passerini. Joint probabilistic-logical refinement of multiple protein feature predictors. *BMC Bioinformatics*, 2013 [Accepted for publication].