COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 126 (2016) 20–34

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Cross-organism learning method to discover new gene functionalities

Giacomo Domeniconi[a,*], Marco Masseroli[b], Gianluca Moro[a], Pietro Pinoli[b]

[a] DISI, Università degli Studi di Bologna, Via Venezia 52, 47521 Cesena, Italy
[b] DEIB, Politecnico di Milano, Piazza L. Da Vinci 32, 20133 Milan, Italy

ARTICLE INFO

Article history:
Received 4 June 2015
Received in revised form
16 November 2015
Accepted 8 December 2015

Keywords:
Biomolecular annotation prediction
Knowledge discovery
Data representation
Discrete matrix completion
Transfer learning
Gene ontology

ABSTRACT

Background: Knowledge of gene and protein functions is paramount for the understanding of physiological and pathological biological processes, as well as in the development of new drugs and therapies. Analyses for biomedical knowledge discovery greatly benefit from the availability of gene and protein functional feature descriptions expressed through controlled terminologies and ontologies, i.e., of gene and protein biomedical controlled annotations. In the last years, several databases of such annotations have become available; yet, these valuable annotations are incomplete, include errors and only some of them represent highly reliable human curated information. Computational techniques able to reliably predict new gene or protein annotations with an associated likelihood value are thus paramount.

Methods: Here, we propose a novel cross-organisms learning approach to reliably predict new functionalities for the genes of an organism based on the known controlled annotations of the genes of another, evolutionarily related and better studied, organism. We leverage a new representation of the annotation discovery problem and a random perturbation of the available controlled annotations to allow the application of supervised algorithms to predict with good accuracy unknown gene annotations. Taking advantage of the numerous gene annotations available for a well-studied organism, our cross-organisms learning method creates and trains better prediction models, which can then be applied to predict new gene annotations of a target organism.

Results: We tested and compared our method with the equivalent single organism approach on different gene annotation datasets of five evolutionarily related organisms (*Homo sapiens*, *Mus musculus*, *Bos taurus*, *Gallus gallus* and *Dictyostelium discoideum*). Results show both the usefulness of the perturbation method of available annotations for better prediction model training and a great improvement of the cross-organism models with respect to the single-organism ones, without influence of the evolutionary distance between the considered organisms. The generated ranked lists of reliably predicted annotations, which describe novel gene functionalities and have an associated likelihood value, are very valuable both to complement available annotations, for better coverage in biomedical knowledge discovery analyses, and to quicken the annotation curation process, by focusing it on the prioritized novel annotations predicted.

* *Corresponding author*. Tel.: +39 3495486309.
E-mail addresses: giacomo.domeniconi@unibo.it (G. Domeniconi), masseroli@elet.polimi.it (M. Masseroli), gianluca.moro@unibo.it (G. Moro), pinoli@elet.polimi.it (P. Pinoli) .

http://dx.doi.org/10.1016/j.cmpb.2015.12.002
0169-2607/

1. Introduction

High-throughput biotechnological methods are generating at increasing rates a progressively high amount of genomic and proteomic data about a growing number of different organisms [1–3]; these mostly regard viruses and bacteria, but also many eukaryotic organisms. They are studied mainly to better understand human patho-physiology or improve food and agricultural products; towards these goals, discovering and comprehension of biological functions of genes and proteins within an organism is paramount.

Besides biological experiments, which are expensive, time-consuming and not always possible, several computational approaches have been proposed to discover gene or protein functions through the processing of the literature or of different types of experimental data; the latter ones include biomolecular sequences, gene expressions, protein interactions and phylogenetic profiles, processed using different techniques (e.g., sequence homology, network based or data and text mining based methods) [4,5]. Some attempts to consider together multiple types of data, also from different species through comparative genomic approaches, have also been proposed in order to improve results (e.g., [6,7]); they are usually quite complex and computationally demanding.

Among available data, those that describe in structured form existing knowledge about structural and functional properties of biomolecular entities (mainly genes and their protein products) are extremely valuable. They are called *controlled biomolecular annotations* and each of them consist of the association of a biomolecular entity with a controlled term, which defines a structural or functional biological property and is part of a terminology or ontology; such association states that the biomolecular entity has the property that the controlled term defines.

Several biomedical terminologies and ontologies exist [8,9]; among them the Gene Ontology (GO) is the most considerable one [10]. It is composed of three sub-ontologies, overall including more than 40,000 concepts, which characterize species-independent Biological Processes (BP), Molecular Functions (MF) and Cellular Components (CC). These concepts are described through controlled terms and are hierarchically related, mainly through *IS_A* or *PART_OF* relationships, within a Directed Acyclic Graph (DAG), designed to capture orthogonal features of genes and proteins. In the GO DAG, each node represents a GO term (i.e., a concept) and each directed edge from a node *A* to a node *B* represents a relationship existing from a child term *A* to its parent term *B*.

Controlled biomolecular annotations are profitably leveraged for biomedical interpretation of biomolecular test results, extraction of novel information useful to formulate and validate biological hypotheses, and also the discovery of new knowledge to classify biomedical literature [11]. They are particularly valuable for high-throughput and computationally intensive bioinformatics analyses. Several computational tools take advantage of these annotations, such as those for *annotation enrichment analysis* (e.g., [12–15]) or *semantic similarity analysis* [16–20] of genes and proteins; they strongly rely on coverage and quality of the available controlled annotations. However, existing controlled biomolecular annotations are accurate only in part, contain errors (particularly those only derived computationally, without human curator supervision) and are incomplete; several biological properties and functions of genes and gene protein products are still to be discovered, especially for recently studied organisms. Furthermore, the available annotations are largely derived computationally, and often they do not have an associated significance level; only a few of them are reviewed by human curators and represent highly reliable information. Annotation curation is crucial for annotation quality; yet, the curation process is very time-consuming. To help and accelerate it, availability of prioritized lists of computationally predicted annotations is greatly effective.

In this scenario, computational techniques able to reliably predict new biomolecular annotations with an associated likelihood value are paramount; towards this aim several different methods have been proposed. King et al. [21] suggested the use of *decision trees* and *Bayesian networks* to predict annotations by learning patterns from available annotations. Tao et al. [22] proposed a *k-nearest neighbour* (k-NN) classifier to associate a gene with new annotations common among its functionally nearest neighbour genes, where functional distance between genes is computed according to the semantic similarity of the GO terms that annotate the genes. *Hidden Markov Model* (HMM) techniques were also used to model gene function evolution [23] or predict gene function from sequential gene expression data [24]. *Support Vector Machine* (SVM) classifiers are also common in gene function prediction. Minneci and colleagues [25] leveraged them to predict GO annotations for several eukaryotic protein sequences, whereas Mitsakakis et al. [26] used them to predict potential functions for previously un-annotated *Drosophila melanogaster* genes, through the analysis of a large dataset from gene expression microarray experiments.

Also biological *network analysis* is often used to predict gene functions. Warde-Farley and colleagues [27] developed GeneMANIA, a server for gene function prediction, where query gene sets are represented as networks with an associated weight, which is based on how well connected the genes in the query set are to each other compared with their connectivity to non-query genes. Differently, Li et al. [28] used a kernel-based learning method and proposed a labeled graph kernel which is able to predict functions of individual genes on the basis of the function distributions in their associated gene interaction networks.

Using a latent semantic approach and basic linear algebra, Khatri and colleagues [29,30] proposed a prediction algorithm based on the *Singular Value Decomposition* (SVD) method of the gene-to-term annotation matrix; this is implicitly based on the count of co-occurrences between pairs of annotation terms in the available annotation dataset. Masseroli and colleagues enhanced this algorithm [31], by including gene clustering based on gene functional similarity computed on the GO annotations of the genes; then, they further extended it by automatically choosing the best SVD truncation level, according to the evaluated dataset [32]. The SVD has also been used with annotation *weighting schemes*, built upon the gene and term frequencies [30,33,34]. Based on simple matrix