Novelty Indicator for Enhanced Prioritization of Predicted Gene Ontology Annotations

Davide Chicco, Fernando Palluzzi, and Marco Masseroli

Abstract—Biomolecular controlled annotations have become pivotal in computational biology, because they allow scientists to analyze large amounts of biological data to better understand test results, and to infer new knowledge. Yet, biomolecular annotation databases are incomplete by definition, like our knowledge of biology, and might contain errors and inconsistent information. In this context, machine-learning algorithms able to predict and prioritize new annotations are both effective and efficient, especially if compared with time-consuming trials of biological validation. To limit the possibility that these techniques predict obvious and trivial high-level features, and to help prioritizing their results, we introduce a new element that can improve accuracy and relevance of the results of an annotation prediction and prioritization pipeline. We propose a novelty indicator able to state the level of "originality" of the annotations predicted for a specific gene to Gene Ontology (GO) terms. This indicator, joint with our previously introduced prediction steps, helps prioritizing the most novel interesting annotations predicted. We performed an accurate biological functional analysis of the prioritized annotations predicted with high accuracy by our indicator and previously proposed methods. The relevance of our biological findings proves effectiveness and trustworthiness of our indicator and of its prioritization of predicted annotations.

Index Terms—biomolecular annotation, prioritized gene annotation, novelty indicator, semantic similarity, Gene Ontology, gene function, functional analysis

1 INTRODUCTION

Massive growth of genomics data and information [1] [2] is forcing scientists to design new concepts and schemes for computational biology. To better express available biological knowledge, and effectively use it to computationally analyze genomics big data, computational biologists introduced controlled biomolecular annotations. A controlled biomolecular annotation is an association between a biomolecular entity (e.g., a gene or a protein) and a controlled term describing one of the biomolecular entity functions. These terms can be part of a flat terminology or of a controlled vocabulary of an ontology, such as the Gene Ontology (GO) [3]; in the latter case semantic hierarchical relationships exist among the controlled terms, so that when a biomolecular entity is annotated to a term, it is also implicitly annotated to all its ancestor terms in the ontology.

Controlled biomolecular annotations are very useful to the scientific community, because they allow scientists to immediately retrieve all the biological function features associated with a specific gene, or vice versa, all genes with a specific function. For example, the statement "the human gene RARA is involved in the molecular function of retinoid acid binding" can be easily expressed with the association between the retinoic acid receptor, alpha (RARA) human gene (identified by the Entrez Gene ID 5914) and the retinoid acid binding term (identified by the ID GO:0001972 of the Gene Ontology). The pairing <RARA, retinoid acid binding> is a typical controlled biomolecular annotation.

Biomolecular annotation databases can be effectively exploited by scientists and researchers to support the understanding of biomolecular test results and the comprehensions of new hypotheses in biology. Many computational tools (e.g., GFINDER [4], FatiGO [6], DAVID [7], QuickGO [8] and others [9]) are available to take advantage of these data resources. Albeit very useful and effective, biomolecular annotation databases also have some important flaws that scientists have to face [10]. First, they are incomplete by definition, since our knowledge of biology is incomplete. Second, they might contain several errors, because only a small percentage of these annotations are supervised by human curators. Third, since different laboratories around the world might work on the same genes or proteins and reach different discoveries, annotation databases might contain inconsistent or ambiguous information about the same genes.

In this context, a key role is played by computational techniques, based upon machine-learning and data-mining algorithms, which are able to predict new biomolecular annotations and generate prioritized lists of them. In the past, we developed several computational intelligence algorithms for this goal.
All these methods are effective in predicting likely Gene Ontology annotations, but they all share a common flaw: most of the gene-function relationships that obtain the highest prediction values often regard obvious high-level functional features (e.g., cellular process). This issue makes such methods predict annotations which are trustworthy, but also often quite trivial and self-evident, and thus not particularly useful for biological discoveries.

To address this problem, we propose an additional layer to the annotation prediction pipeline: a novelty indicator able to state the “originality” of the annotations predicted for a gene, and thus helping their prioritization. Furthermore, we show that taking into account this indicator as an additional step in our prediction pipeline leads to the prioritization and accurate prediction of GO annotations having strong biological relevance, as proved through an accurate functional analysis.

We organize the rest of this paper as follows. After this Introduction, we describe some previous work related to our novelty indicator in Section 2. We then describe the datasets that we used for testing, the implemented novelty indicator, and the main results of its application to predicted biomolecular annotations in Section 3. In the second part of the paper, we report the prioritized predicted annotations obtained, and the techniques we used to find and select them in Section 4.1. Then, in Section 4.2 we describe the functional analysis that we made to highlight the biological relevance of the prioritized predicted annotations obtained. Finally, we illustrate the main conclusions and possible future developments in Section 5.

2 Related works

In the past twenty years, scientists developed several measures to state the level of semantic similarity between two genes or proteins. In a survey by Pesquita et al. [22], the authors comprehensively described the main semantic similarity measures used in the biomedical ontology domain, providing also some examples of implementations and applications, and a complete comparison. In particular, they described the Jiang [23], Lin [24], and Resnik [25] scores, which are able to measure the semantic similarity between genes (or gene products) taking advantage of the structure of the analyzed ontology. All of these are information-theoretic approaches based upon the concept of lowest common ancestor (LCA) between two ontology terms, where the LCA is the closest ancestor node that the two terms have in common in their ontology (that is their lowest shared ancestor in the ontology tree structure). These information-theoretic measures were shown to be significantly more robust than other scores, especially with respect to the node density variability in different branches of the ontology.

Other scientists then invented more complex measures, which tried to integrate this information-theoretic knowledge with other available biological information. Lord and colleagues [26] introduced a protein similarity approach that combines protein sequence similarity (computed through bioinformatics tools such as BLAST [27]) and semantic similarity based upon protein GO annotations (computed through classical measures such as the Resnik one [25]). Their measure generates interesting results, but does not consider the position of the GO terms in the ontology tree, and is very bounded to the GO Molecular Function sub-ontology. Conversely, Speer et al. [28] proposed a similarity measure that takes advantage of a clustering technique for the partition of genes according to their GO biological functions. This technique leads to good results, but its limitation is the clustering distance choice: the authors showed that the selection of slightly different clustering distances might lead to very different similarity results.

Starting from the two similarity measures by Lord and colleagues [26] and Speer et al. [28], Schlicker et al. [29] properly modified their scores, and developed a new indicator, named GOscoreBM, which is able to take advantage of both the structural position of the terms in the analyzed ontology tree and their semantic similarity score computed through the Resnik measure. Due to its completeness of information, we decided to take advantage of this Schlicker score as novelty indicator within our work on prediction and prioritization of Gene Ontology annotations.

It is also worth mentioning QuickGO [8], a GO browser with visualization functionalities able to show a GO directed acyclic graph (DAG), i.e., the tree structure of a sub-part of the GO induced by the hierarchical relationships existing between GO terms. Its web interface provides an easy-to-use DAG of the ancestor terms of a GO term, whose ID or name is specified by the user. However, with QuickGO (and with any of the GO visualization tools currently available) it is not possible to differentiate existing annotation terms from new predicted annotation ones.

To detect the “novelty” of predicted gene annotations, we implemented a statistical measure able to compare the ontological trees of the annotation terms of a gene before and after the annotation prediction, and to evaluate the dissimilarity of the two ontological trees. In the next section we introduce this novelty indicator which is based on Schlicker et al. work [29].

3 Methods

In this section we describe the annotation prediction and prioritization pipeline and the datasets that we used for our tests, as well as the novelty indicator that we applied to enhance the prioritization of the predicted annotations, and its main application results.
3.1 Prediction pipeline and datasets

We used the annotation prediction and prioritization pipeline we previously described in [21], which includes the prediction methods: truncated Singular Value Decomposition (tSVD), Semantically Improved tSVD with gene clustering (SIM1), and Semantically Improved tSVD with gene clustering and feature term similarity weights (SIM2) [12] [13]. Figure 1 shows the used computational pipeline and its extension with the novelty indicator proposed in this paper.

We ran our prediction tests on the datasets of Gene Ontology annotations of Homo sapiens genes available in the Genomic and Proteomic Data Warehouse (GPDW) [30] [31], an integrated data resource publicly and freely available from Politecnico di Milano at http://www.bioinformatics.deib.polimi.it/GPKB/ which includes multiple versions of annotation datasets. We applied our prediction pipeline on the annotations of the July 2009 GPDW version and then validated the predicted annotations by looking for them in the March 2013 GPDW version [32]. Despite the March 2013 not being the most updated GPDW version, we used it because it is one of the most stable and accurate versions recently delivered [30].

We chose the Homo sapiens gene annotations to the three GO sub-ontologies (Biological Process, Molecular Function, Cellular Component) because they include representative numbers of genes and GO terms. We excluded all the annotations having evidence code equal to IEA (Inferred from Electronic Annotation) or ND (No biological Data available) from the input dataset in order to base our prediction only on the most reliable annotations available. Conversely, we made no evidence code distinction when considering the annotations from the more recent GPDW version, i.e., we considered all available annotations (including the computational ones) to validate our predicted annotations. We are aware of the importance of the evidence code information, and we plan to use it as an additional selection layer to our pipeline in the future.

In the July 2009 analyzed dataset, the Homo sapiens gene GO annotations had the following quantitative characteristics: for the Biological Process (BP) sub-ontology: 7,902 genes, 3,528 GO terms, and 21,048 annotations; for the Molecular Function (MF) sub-ontology: 8,590 genes, 2,057 GO terms, and 15,467 annotations; for the Cellular Component (CC) sub-ontology: 7,868 genes, 684 GO terms, and 14,341 annotations.

3.2 Novelty indicator

Schlicker and colleagues originally introduced the GOscoreBM semantic similarity measure to evaluate the similarity between two genes based upon their GO annotations [29]. Conversely, we propose to take advantage of this measure as a statistical indicator of the novelty of ontological annotations predicted for a gene, by using it to compare the DAG of the gene annotation terms before and after the prediction. The rationale of our approach is the following. For each analyzed gene, we have a tree of its ontological annotations before the prediction, and another tree of its annotations after the prediction, in which the new predicted annotations have been added. We can quantify the semantic difference between the before-prediction tree and the after-prediction tree, as evaluated by the GOscoreBM measure, and consider it as the level of “novelty” brought by the prediction made. Therefore, we compute this semantic difference...
through the GOScoreBM, defined as follows.

Given two genes \( p \) and \( q \), let us denote with \( GO^p \) the set of all GO terms annotated to the gene \( p \), and with \( GO^q \) the set of all GO terms annotated to the gene \( q \). The general idea of the GOScoreBM is to build a matrix \( S \) with the pairwise semantic similarity measures between each term in the first set \( GO^p \) and each term in the second set \( GO^q \); then, to consider the maximum value of each row of the matrix \( S \) and to compute the average of these values, doing likewise also for the matrix columns.

We build the \( S \) matrix as follows: with \( i = 1, \ldots, N \) rows and \( j = 1, \ldots, M \) columns, each matrix element \( s_{ij} \) is computed as the Resnik similarity measure \([25]\) between the \( i_{th} \) element in the \( GO^p \) set and the \( j_{th} \) element in the \( GO^q \) set. This way, we compute the semantic similarity between each possible pair of terms in the considered GO annotation sets, i.e.:

\[
\forall i \in \{1, \ldots, N\}, \forall j \in \{1, \ldots, M\}:
    s_{ij} = \text{ResnikSimilarity}(GO^p_i, GO^q_j)
\]

Based on this \( S \in \mathbb{R}^{N \times M} \) matrix, two operators are defined: \( \text{rowScore} \), as the average of the row maximum values of the \( S \) matrix, and \( \text{columnScore} \), as the average of the column maximum values of \( S \):

\[
\forall i \in \{1, \ldots, N\}, \forall j \in \{1, \ldots, M\}:
    \text{rowMaxScore}_i = \max(s_{ij})
\]

\[\sum \text{rowMaxScore}_i\]

\[
\forall j \in \{1, \ldots, M\}, \forall i \in \{1, \ldots, N\}:
    \text{columnMaxScore}_j = \max(s_{ij})
\]

\[\sum \text{columnMaxScore}_j\]

\[
\text{rowScore} = \frac{\text{rowMaximaSum}}{N}
\]

\[
\text{columnScore} = \frac{\text{columnMaximaSum}}{M}
\]

Indeed, \( \text{rowScore} \) and \( \text{columnScore} \) serve to find, for every GO term annotated to a gene, the best-matching GO term annotated to the other gene estimate the average maximum similarity between the annotations of the two given genes \( p \) and \( q \) \([33]\). Then, the GOScoreBM measure for the two genes is defined as:

\[
\text{GOScoreBM}(p, q) = \max(\text{rowScore}(p, q), \text{columnScore}(p, q))
\]

Schlicker and colleagues introduced this selection between \( \text{rowScore} \) and \( \text{columnScore} \) to reward the subset of most similar terms of the annotations of the two genes, subset of terms for each gene, by assigning more weight to the most similar terms and less weight to the less similar ones and demonstrated that this provides a good evaluation of the semantic similarity of the two genes, i.e., of the similarity of their annotation trees \([29]\).

We use this GOScoreBM score to measure the level of “novelty” of the annotations predicted for a gene by comparing the set of GO terms annotated to the gene before the prediction and the set of terms annotated to it after the prediction; \( \text{BM} \) \( g_{after} \). The more the two sets (i.e., the two annotation trees of the gene before and after the prediction) are different, the lower the GOScoreBM is. Therefore, the GOScoreBM measure allows selecting, among all the predicted annotations, the “interesting” non-obvious ones; in fact, these annotations regard terms in medium-upper levels of the annotation ontology which, together with their ancestor terms in the lower levels of the ontology, induce relevant changes in a gene annotation tree when predicted.

Among all the most common semantic similarity measures available between two terms of an ontology (Jiang \([23]\), Lin \([24]\), and Resnik \([25]\)), we decided to use the Resnik one because it is considered the most efficient score in correlating gene function similarities \([34, 35]\). The fact that the Resnik similarity measure has no upper bound, inducing the GOScoreBM score that uses it to have no predefined upper bound, does not influence our application, since we look for low values of the score, which we heuristically defined as GOScoreBM \( < 1 \).

The main advantage of introducing this novelty indicator is to help the computational machinery to select automatically interesting non-obvious annotations among all the predicted ones. As previously explained \([20]\), this is a limit of any algorithm for the prediction of ontological annotations, i.e., often most of the predicted gene-function relationships are rather obvious, high-level, trivial descriptive features, such as the general cell cycle feature. To address this issue, we decided to use this novelty indicator as an additional tool to help computationally predicting and prioritizing annotations not only very likely to be correct, but also deemed novel and interesting, since quite different from those known before the prediction.

### 3.3 Novelty indicator tests

We tested the use of the GOScoreBM measure as a novelty indicator on all the GO annotations predicted for the Homo sapiens genes, based upon the gene GO annotations available in the GPDW dataset of July 2009 and using the tSVD, SIM1 and SIM2 methods \([12]\). The GOScoreBM measure resulted in concordance with the visual evaluation of the gene annotations performed by an expert.

As a general example, in Figure 2 we show the DAG of the GO Biological Process terms annotated to the human protein phosphatase methylesterase 1 (PPME1) gene. The annotation prediction for this gene led to a very low (good) indicator value (GOScoreBM = 0.087).
We used the described novelty indicator as the last step of our computational pipeline for annotation prediction [21], to create a final list of prioritized annotations very likely to be correctly predicted and interestingly novel. In Section 4.1 we describe these annotations that we obtained, and in Section 4.2 we report their thorough functional analysis that we performed to demonstrate their correctness and relevance.

We ran the prediction tests on Homo sapiens gene GO annotations of the previously-mentioned GPDW dataset of July 2009, by using the tSVD, SIM1 and SIM2 methods (Table 1).

On the basis of the prediction algorithms and novelty indicator described or referenced in the previous section, we can prioritize the most likely predicted annotations which respect strict accuracy conditions. In Table 2 we list such GO Biological Process predicted annotations obtained for the considered

---

**Table 1**

Quantitative characteristics of the Homo sapiens GO annotations predicted in the tests. tSVD, SIM1 and SIM2 are prediction algorithms [12]. Cellular Component, Molecular Function and Biological Process are the sub-ontologies of the Gene Ontology.

<table>
<thead>
<tr>
<th>method</th>
<th>Cellular Component</th>
<th>Molecular Function</th>
<th>Biological Process</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>tSVD</td>
<td>8</td>
<td>64</td>
<td>64</td>
<td>261</td>
</tr>
<tr>
<td>SIM1</td>
<td>8</td>
<td>35</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>SIM2</td>
<td>8</td>
<td>14</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>261</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
dataset, which are not found in the more updated GPDW version considered (of March 2013) and satisfy the following conditions:

[i] Predicted by one or more of the used methods (tSVD, SIM1, SIM2) [12];

[ii] Ranked within the top 10% of the predicted annotations (“top 10%” column in Table 2), according to the annotation likelihood calculated by the prediction method, which means being one of the most likely predicted annotations; than our fixed threshold $\omega = 2/3 = 66.67\%$ (this condition is true for all our predicted gene annotations and therefore it is not reported in the table) [32];

[iii] Regarding a gene with more than 50% of the predicted annotations found confirmed in the more updated GPDW database version considered (“conf. > 50%” column in Table 2);

[iv] Regarding a GO term with all the parent terms already annotated to the same gene in the considered GPDW database version, or with at least one parent term predicted annotated to the same gene and found confirmed in the more updated GPDW database version considered (“pred. conf.” column in Table 2), as in the “prioritization rule” introduced in [21];

[v] Low novelty indicator value (“GOScoreBM < 1” column in Table 2), which indicates a novel interesting prediction for the gene (this condition is true for all the genes in Table 2); we heuristically evaluate novel interesting enough annotations to be considered only the predictions leading to novelty indicator values lower than 1.

We report all these prioritization conditions in Figure [1] where we label them with their corresponding numerals (from [i] to [v]). Taking advantage of our computational pipeline and of these conditions, we can predict and prioritize reliable accurate annotations which are also biologically relevant and interesting, because they were selected by the novelty indicator score.

4.1 Prioritized predicted annotations

In Table 2 we list the novel GO Biological Process annotations predicted for the Homo sapiens genes, based upon the old GPDW version considered (of July 2009), not found reported in the more updated GPDW database version considered (of March 2013), and that satisfy both the novelty indicator condition and at least other three of the above prioritization conditions (sorted by decreasing number of conditions satisfied). Among all the 261 predicted gene GO annotations reported in Table 1, only the 7 listed in Table 2 satisfy at least four items of the [i]-[v] accuracy conditions; thus, they represent only the top 2.68% of all the gene-function relationships predicted by our algorithms. They would not be obtained as such without using the proposed novelty indicator.

The novelty indicator condition states the relevance of the prediction: without it, our pipeline would select mainly obvious and trivial gene-function relationships, limiting its ability to lead to significant biological discoveries. The added value of the novelty indicator to our pipeline consists in constraining the prioritized predicted annotations to be “novel enough” to raise the interest of the biologists’ community, as we address in the functional analysis in Section 4.2. Our novelty indicator defines as biologically interesting relevant an annotation that is part of a prediction which introduced many relevant nodes to novel predicted annotations that introduce many nodes in the annotation tree of a gene, i.e., whose terms belong to high, specific levels of the annotation ontology; this induces a high semantic difference between the before-prediction tree and the after-prediction tree of the gene, thus a low GOScoreBM.

The prioritized predicted annotation list (Table 2) is the final biological relevance we can computationally provide to physicians and biologists regarding their experiments about human genes. It is a concrete application of our methods, that we hope might improve and quicken the discovery of new cures, new therapies, and new knowledge about gene functions. sorted by number of conditions satisfied, from the ones that satisfy more conditions, to the ones that satisfy less conditions.

By observing Table 2, we can notice that the annotation "<PPME1, organelle organization>" was predicted by all three prediction methods used, has a likelihood score that ranks it in the top 10% of all the predicted annotations, and refers to a GO term with at least one parent term predicted annotated to the same gene and found confirmed in the updated version of the GPDW database.

Two annotations predicted for the CHST14 gene (<CHST14, chondroitin sulfate proteoglycan biosynthetic process> and <CHST14, biopolymer biosynthetic process>) are suggested by all the thee methods used (tSVD, SIM1 and SIM2), regard a gene with more than a half of the predicted annotations found confirmed in the more updated version considered of the GPDW database, and regard GO terms with at least one parent term predicted annotated to the same gene and found confirmed in the more updated version considered of the GPDW database. Despite that, their prediction likelihood score does not rank them in the top 10% of the annotations predicted by all the three methods.

Also the annotation <CHST14, dermatan sulfate proteoglycan biosynthetic process> was predicted by the three methods considered, and refers to a gene with more than half of the predicted annotations found confirmed in the more updated version considered of the GPDW database.

The annotation <CPA2, proteolysis involved in cellular protein catabolic process> was predicted only by the
**TABLE 2**

List of the top prioritized gene annotations to the GO Biological Process (BP) sub-ontology predicted by our methods and not found in the more updated version of the GPDW database considered. The "# conditions" column states how many conditions of this table were satisfied by the specific gene annotation predicted. The "predicted by tSVD, SIM1, SIM2" columns state which method(s) predicted the annotation; the "top 10%" column states if, in the likelihood ranking of the three the considered prediction methods, the annotation position is in the top 10% of all the annotations predicted; the "conf. > 50%" column states if the percentage of all the predicted annotations for the gene that are found confirmed in the more updated version of the GPDW database considered is greater than 50%; the "pred. conf." column states if the predicted annotation term has all the parent terms annotated to the gene in the considered GPDW database version, or it has at least one parent term predicted annotated to the gene that is found confirmed in the more updated version considered of the GPDW database; the "GOscoreBM < 1" column states if the introduced novelty indicator has a low value for the annotated gene, that indicates a novel interesting prediction for the gene.

<table>
<thead>
<tr>
<th># conditions</th>
<th>gene symbol (Entrez Gene ID)</th>
<th>GO term description</th>
<th>predicted by tSVD</th>
<th>SIM1</th>
<th>SIM2</th>
<th>top 10%</th>
<th>conf. &gt; 50%</th>
<th>pred. conf.</th>
<th>GOscoreBM &lt; 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>PPME1 (51400)</td>
<td>Organelle organization. [BP] (GO:0006996)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>6</td>
<td>CHST14 (113189)</td>
<td>Chondroitin sulfate proteoglycan biosynthetic process. [BP] (GO:0050650)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>6</td>
<td>CHST14 (113189)</td>
<td>Biopolymer biosynthetic process. [BP] (GO:0043284)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>5</td>
<td>CHST14 (113189)</td>
<td>Dermatan sulfate proteoglycan biosynthetic process. [BP] (GO:0050651)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>5</td>
<td>CPA2 (1358)</td>
<td>Proteolysis involved in cellular protein catabolic process. [BP] (GO:0051603)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>PPME1 (51400)</td>
<td>Chromosome organization. [BP] (GO:0006996)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>CNOT2 (4848)</td>
<td>Positive regulation of cellular metabolic process. [BP] (GO:0031325)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

SIM1 and SIM2 methods, but it refers to a gene with more than half of the predicted annotations found confirmed in the more updated version considered of the GPDW database, and validated refers to involves a GO term with at least one parent term predicted annotated to the same gene and found confirmed in the more updated version considered of the GPDW database.

Another annotation, <PPME1, chromosome organization>, was predicted by the tSVD and SIM1 methods only, but its likelihood score ranks it within the top 10% of the predicted annotations.

Finally, in this list of the prioritized most likely annotations predicted, the gene annotation <CNOT2, positive regulation of cellular metabolic process> was predicted only by the tSVD method, but its likelihood score ranks it in the top 10% of the predicted annotations, and it refers to a GO term with at least one parent term predicted annotated to the same gene and found confirmed in the more updated version considered of the GPDW database. Conversely from the other prioritized annotations which involve quite specific GO terms, this annotation refers to a quite general function and low level GO term; despite that, it is relevant since it has a clear supporting evidence: CNOT2 is part of the CCR4-NOT transcription complex (comprising CNOT, TOB1, RQCD1 genes; see Table [3]), a highly conserved machinery with a general role in controlling mRNA metabolism, including mRNA degradation and miRNA-induced silencing.
transcription initiation and elongation, ubiquitination, and post-transcriptional regulation [36, 39]. Members of the CCR4-NOT complex interact with components of the proteasome (including PSMA proteins and ubiquities UBE2D1, UBE2E1 and UBC) and with poly-A binding proteins (including PABPC1 and PAIP1).

As mentioned, 261 gene GO annotations were predicted, but only 7 of them (2.68%) satisfied both the novelty indicator condition and at least three of the other six conditions reported in Table 2.

4.2 Functional analysis

The application of the accuracy controls previously described, which include our novelty indicator, prioritized the seven GO Biological Process (BP) annotations predicted for the four Homo sapiens genes listed in Table 2. To assess the correctness and biological relevance (i.e., interestingness) of these novel annotations predicted, we evaluated them using a network-based functional validation procedure, followed by a cross-check against the KEGG pathway database [39]. Such procedure allows supporting a predicted annotation when the involved gene is closely related, in a gene network, to other genes that are known to be annotated to the same term of the predicted annotation.

The performed biological assessment highlighted the importance of our prediction pipeline and accuracy controls in reliably predicting and prioritizing new interesting gene annotations, therefore improving current biological knowledge. As mentioned, the addition of the novelty indicator presented in this paper allowed our computational prediction pipeline to avoid selecting obvious low-level descriptive features, and to finally prioritize annotations deemed novel enough to raise the biologists’ community interest.

For each gene in Table 2, we retrieved a network of interacting neighbor genes from the STRING database [40]. We chose STRING for three reasons: (i) to take advantage of the different data types integrated in STRING, including experimental assays, inferred interactions, and text mining information; (ii) to base our evaluation on a comprehensive annotation repository, integrating knowledge from several independent sources; and (iii) to rely on a combined score for interaction filtering, which STRING provides. Furthermore, the STRING database allowed us to retrieve predicted functional partner (PFP) genes of our genes, according to the STRING annotation, in order to improve the analysis. To include only strong associations, we considered a relationship only if its STRING combined score is above 0.6.

The first step of our validation procedure is the gene network expansion of each considered gene $x$, which is annotated to a set of GO terms $\{t_i\}$. Let us define the gene network expansion depth as the minimum number of steps to walk from the farthest neighbor gene to the gene $x$, or to one of its PFP genes. The latter ones, provided by STRING, are used to improve our search for functionally related genes that could be enriched for the predicted annotation terms. The interactome obtained after the expansion procedure contains genes annotated to at least one of the $t_i$ terms; some of these genes can be PFP genes. There can be also non-PFP genes in the interactome, which can be annotated to the $t_i$ terms; we define them as association-supporting genes (ASGs), since they are genes interacting with a considered gene $x$ and annotated to at least a $t_i$ term. We heuristically found that the optimal depth for the expansion of our predicted annotations is 3; this allowed us to find evidence, at least indirect, for six out of the seven predicted and prioritized annotations. The only exception was the predicted annotation of the human carbohydrate (N-acetylgalactosamine 4-0) sulfotransferase 14 (CHST14) gene to the biopolymer biosynthetic process (GO:0043284), which could not be validated at any reasonable depth ($\leq 10$).

Furthermore, we classified the functional evidence for a predicted annotation to a term $t_x$ in four classes: (i) already confirmed, when available in public databases at the time of writing; (ii) primary evidence, when one or more PFP genes are annotated to $t_x$; (iii) secondary evidence, when one or more non-PFP genes are annotated to $t_x$; (iv) no evidence, when there are no genes supporting the annotation. In many cases, a predicted annotation can be supported both primarily and secondarily.

Table 3 reports the results for gene network expansion followed by GO Biological Process and KEGG enrichment for the prioritized predicted annotations of Table 2.

For every considered gene, each selected annotation was tested using the aforementioned procedure. Out of the 7 prioritized predicted annotations listed in Table 2 only one, <CHST14, dermatan sulfate proteoglycan biosynthetic process (GO:00050651)>, was available (i.e., confirmed) at the time of writing. The other human CHST14 gene annotation to chondroitin sulfate proteoglycan biosynthetic process (GO:00050650), which is not available in public databases, is shared with five PFP genes (i.e., BCAN, NCAN, VCAN, CSPG4, CSPG5) and two ASGs (i.e., Dcn, Igf1) of the CHST14 gene in its expanded network at depth 3 (Figure 3). The role of these genes in inflammation has been largely documented in literature. All the detected genes are involved in mast cell secretion during inflammatory response, characterizing both innate and adaptive immunity. Mast cells are characterized by a large amount of cytoplasmatic secretory granules, containing negatively charged molecules, including heparin and chondroitin sulfate proteoglycans. Furthermore, Igf1 production is known to be stimulated by CD44 induction by hyaluronic acid during macrophage-mediated inflammatory and repair processes [31, 33].
### TABLE 3

Biological assessment of the prioritized novel gene annotations predicted. Results of network expansion followed by Gene Ontology (GO) Biological Process (BP) and KEGG enrichment are shown. Columns PFPs (Predicted Functional Partner genes) and ASGs (Association-Supporting Genes) report the ensemble of genes found supporting the prioritized gene annotations predicted. PFPs are predicted interactors of the annotated gene, according to STRING [40]. ASGs are interacting genes sharing the predicted annotation. The two gene lists overlap when PFPs have the predicted annotation. KEGG enrichment is used as an external source to further support the predicted annotation to the biological process. Enrichment is considered significant if its p-value is less than 0.01.

<table>
<thead>
<tr>
<th>gene</th>
<th>Entrez Gene ID</th>
<th>GO BP</th>
<th>GO ID</th>
<th>biological validation</th>
<th>PFPs</th>
<th>ASGs</th>
<th>KEGG enrichment (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPME1</td>
<td>51400</td>
<td>Organelle organization</td>
<td>GO:00066996</td>
<td>Primary evidence</td>
<td>TJAP1, PPP2CA</td>
<td>AKT1, AXIN1</td>
<td>hsa04530: Tight junction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chromosome organization</td>
<td>GO:0051276</td>
<td>Secondary evidence</td>
<td>PPP2R1A, PPP2R2B</td>
<td>STRIP1</td>
<td>(3.1E-7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PPP2CB, PPP2R2A, PPP2R2A</td>
<td>PPP2CA</td>
<td>hsa04310: Wnt signaling pathway (9.3E-4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PPP4R1L, PPP4C, LCMT1, XRRA1, DNAJB13</td>
<td></td>
<td>hsa04076: Long-term depression (1.7E-3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dermatan sulfate proteoglycan biosynthetic process</td>
<td>GO:0050651</td>
<td>Already confirmed</td>
<td>BCAN, NCAN, VCAN, CSPG4, CSPG5, DCN</td>
<td>BCAN, NCAN, VCAN, CSPG4, CSPG5, DCN</td>
<td>hsa04350: TGF-beta signaling pathway (2.5E-3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chondroitin sulfate proteoglycan biosynthetic process</td>
<td>GO:0050650</td>
<td>Primary evidence</td>
<td>B3GAT1, ZNF469</td>
<td></td>
<td>hsa04014: Oocyte meiosis (4E-3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Biopolymer biosynthetic process</td>
<td>GO:0043284</td>
<td>No evidence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHST14</td>
<td>113189</td>
<td>Chondroitin sulfate proteoglycan biosynthesis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNOT2</td>
<td>4848</td>
<td>Positive regulation of cellular metabolic process</td>
<td>GO:0031325</td>
<td>Primary evidence</td>
<td>CNOT1, CNOT3, CNOT4, CNOT6, CNOT6L, CNOT7, CNOT8, CNOT10, TOB1, RQCD1</td>
<td>CPEB3, PAIP1, CNOT1, CNOT8, TOB1, PABPC1, UBC, UBE2D1, UBE2E1, PSA1, PSA2, PSA3, PSA4, PSA6, PSMC3, PSMC7</td>
<td>hsa03018: RNA degradation (1.3E-8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CPA2</td>
<td>1358</td>
<td>Proteolysis involved in cellular protein catalytic process</td>
<td>GO:0051603</td>
<td>Secondary evidence</td>
<td>CTRB1</td>
<td></td>
<td>hsa03050: Proteasome (2.2E-8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

KEGG enrichment evaluation shows that all these genes are involved in chondroitin sulfate biosynthesis (p-value < 0.001). Furthermore, on the base of the GO Biological Process DAG, chondroitin and dermatan sulfate proteoglycan biosynthesis are tightly related. All this provides further molecular support to the predicted annotation of chondroitin sulfate proteoglycan biosynthetic process to the CHST14 gene through its PFPs and ASGs. Thus, proximity and predicted associations enabled us to validate the predicted annotation from a biological functional point of view; furthermore, our prioritized prediction of this annotation correctly added related knowledge of biological interest, not present in GO nor in KEGG databases, to an existing new gene annotation (<CHST14, dermatan sulfate proteoglycan biosynthetic process>).

We validated the predicted annotation of the human protein phosphatase methylesterase 1 (PPME1) gene...
to organelle organization (GO:0006996) through the 9 PPME1-interacting genes (i.e., ASGs) already annotated to organelle organization, which include the AKT1, AXIN1, STRIP1, PPP2R4, PPP2R2A, PPP2CA, PPP2R1A, PPP2CB and TJAP1 genes (Figure 4); only the last four of these genes are also PFP genes according to STRING, together with the PPP2R1B, PPP4R1L, PPP4C, LCMT1, XRA1 and DNAJB13 genes (Table 3). Please notice that, since organelle organization is the parent term of chromosome organization, all genes annotated to the latter one are also annotated to the former one.

The other human PPME1 predicted annotation, to chromosome organization (GO:00051276), was not found in any of the PPME1 interactors at depth 3. However, adding its first five ASG genes (i.e., AKT1, AXIN1, STRIP1, PPP2R4, PPP2R2A) to its PFP gene set and repeating the analysis at depth 3, also the annotation to chromosome organization (GO:00051276) is found in four distinctly interacting genes (not direct interactors and non-PFPs) of PPME1. These genes are LEF1, TCF7L2, CDKN2A and PTGES3, all involved in cancer and neural development pathways (as confirmed by KEGG over-representation analysis - data not shown), which is in accordance with the function of the AKT1, AXIN1, STRIP1, PPP2R4, PPP2R2A, PPP2CA, PPP2R1A, PPP2CB and TJAP1 genes. This is evidence for the predicted annotation of PPME1 to chromosome organization, since we included possible functional partners of PPME1, even if distantly placed in the interactome, using the two related predicted annotations, i.e., organelle organization and chromosome organization. This suggests that ASGs can be proposed as novel PFP genes, since they show enrichment in specific biological processes (in our case, chromosome organization). In the specific case of PPME1, a nuclear phosphatase, the predicted annotation to chromosome organization is supported by 9 ASGs, including several PP2A phosphatases (PPP2R4, PPP2R2A, PPP2CA, PPP2R1A, PPP2CB).

Deregulation of phosphatases PP2A is a common biomarker of various complex diseases including breast cancer [45] and Alzheimers disease [46]. In these works, it has been reported how the expression of these phosphatases is controlled by cytoskeleton-associated factors and cofactors involved in chromosome and organelle organization, including AXIN1, LCMT1 and STRIP1 (that are in the ASG list of PPME1). All this supports our prioritized prediction of PPME1 to chromosome organization, and shows how a gene can be added to the existing knowledge of a biological process, suggesting a new possible interaction even with distantly related genes.

A clear example of the possibility of proposing ASGs as novel PFP genes also exists for the predicted annotation of the human carboxypeptidase A2 (pancreatic) (CPA2) gene to proteolysis involved in cellular protein catabolic process (GO:0051603). None of the PFP genes of CPA2 were found annotated to the prioritized predicted annotation term. However, after interactome expansion, we detected a gene supporting this predicted annotation: the UBC gene, which...
Fig. 4. Expanded (at depth 3) interaction network of the human \textit{PPME1} gene. Enrichment was found for the annotation to organelle organization (GO:0006996). Dark red nodes support the predicted annotation. Color code of the connections is as in STRING \cite{40} (green: neighborhood; red: gene fusion; blue: co-occurrence; black: co-expression; purple: experiment-based; cyan: database-based; light green: text mining; violet: homology).

The gene \textit{PPME1} encodes for a precursor of Ubiquitin and interacts with \textit{POR} and \textit{ATP6V0D1}, two ASGs of \textit{CPA2}. \textit{CPA2} is a pancreatic carboxypeptidase involved in insulin metabolism \cite{47}.

In our prioritized prediction, \textit{CPA2} has been annotated with proteolytic activity. The ASGs we found for this gene share, by definition, the same annotations. These genes are also involved in insulin metabolism, diabetes mellitus, and immune response. Although \textit{CPA2} proteolytic activity is documented since long time \cite{48,49}, the relationship with insulin metabolism has not been clarified yet. In a recent work \cite{50}, it has been shown how high expression levels of genes (including \textit{CPA1}, \textit{CPA2} and \textit{CTRB1}) involved in insulin sensitivity, erythropoiesis, hemangioblast generation and cellular redox control were evident in spleens of cured mice, which indicates their possible contribution to protection against autoimmune type 1 diabetes mellitus. Thus, in this case, our prioritized prediction is supported by the literature, and prompted us to recover also additional information about other related functions connecting \textit{CPA2} to its ASGs (i.e., insulin metabolism and immune response).

Besides assessing the validity and biological relevance of our prioritized predicted gene GO annotations, this functional analysis procedure provides a way of finding new functional partners of our considered genes, using predicted knowledge. In general, our results show how this validation procedure can be used to find novel genes involved in a biological process, being functional partners of genes already known to be involved in the process.

5 Conclusions
Biomolecular annotations are pivotal concepts in computational biology, but unfortunately they contain errors and are always incomplete by definition, since incomplete is our knowledge of biology. Thus, machine-learning and data-mining algorithms able to reliably predict them can be effective tools to suggest new gene functions to biologists and biomedical researchers.

In the past, we developed and applied several annotation prediction methods and a prioritization rule able to provide trustworthy annotations. Here, we extended our previous projects by introducing a novelty indicator able to state the level of “originality” of the predicted Gene Ontology annotations of a gene. We showed that it helps to reliably prioritize the predicted gene annotations, and select relevant annotations that would not been prioritized otherwise.

We performed an in-depth functional analysis of the prioritized predicted annotations obtained, which highlights the biological relevance of the most promising biomolecular annotations that our methods predicted and the introduced novelty indicator helped prioritizing. Results showed novel interesting biological aspects that can be leveraged by biologists and biomedical scientists.

In the future, we plan to apply the novelty indicator to biomolecular annotations predicted through other
computational methods (such as latent Dirichlet allocation, probabilistic latent semantic analysis, or deep autoencoder neural networks) based upon the most recent Gene Ontology annotation dataset available. For future enhanced functional analysis, we consider to use public domain databases such as Reactome [51] in addition to KEGG, used here. We also aim to integrate our computational pipeline into the online Bio Search Computing framework [52, 53], publicly available through the Internet.

ACKNOWLEDGMENTS

Authors thank Eric G. Roberts (Princess Margaret Cancer Centre) for the English proof-reading of this manuscript.

REFERENCES


This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBBB.2017.2695459, IEEE/ACM Transactions on Computational Biology and Bioinformatics.


Marco Masseroli received his Bachelor and Master of Science degree in computer science at Università di Genova (Genoa, Italy), respectively in 2007 and 2010. He then started the PhD program in computer engineering at Politecnico di Milano (Milan, Italy), where he graduated in Spring 2014. He also spent a semester as visiting research scholar at University of California Irvine (Irvine, California, USA). Since September 2014, he has been a post-doctoral researcher at the Princess Margaret Cancer Centre and guest at University of Toronto (Toronto, Ontario, Canada). His research topics focus upon mainly machine learning algorithms applied to bioinformatics.

Davide Chicco obtained his Bachelor of Science in biology at Sapienza Università di Roma (Rome, Italy) in 2008, a postgraduate degree in bioinformatics at Università di Siena (Siena, Italy) in 2010, and then his Master of Science in bioinformatics at Università di Bologna (Bologna, Italy) and Universitat Pompeu Fabra (Barcelona, Spain) in 2012. Since November 2012, he has been a PhD student at the Dipartimento di Elettrofisica, Informazione e Bioingegneria of Politecnico di Milano (Milan, Italy). His research interests include high throughput sequencing data integration and analysis, applied to cancer genomics. He is currently involved in joint projects at Istituto Europeo di Oncologia - IEO (Milan, Italy) for the characterization, classification and discovery of structural chromatin rearrangements and DNA damage in leukemia.

Fernando Palluzzi obtained his Bachelor of Science in biology at Sapienza Università di Roma (Rome, Italy) in 2008, a postgraduate degree in bioinformatics at Università di Siena (Siena, Italy) in 2010, and then his Master of Science in bioinformatics at Università di Bologna (Bologna, Italy) and Universitat Pompeu Fabra (Barcelona, Spain) in 2012. Since November 2012, he has been a PhD student at the Dipartimento di Elettrofisica, Informazione e Bioingegneria of Politecnico di Milano (Milan, Italy). His research interests include high throughput sequencing data integration and analysis, applied to cancer genomics. He is currently involved in joint projects at Istituto Europeo di Oncologia - IEO (Milan, Italy) for the characterization, classification and discovery of structural chromatin rearrangements and DNA damage in leukemia.