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 their test results, and to infer new knowledge. Yet, biomolecular annotation databases are incomplete by definition, like our knowledge of biology, and may contain errors and inconsistent information. In this context, machine-learning algorithms able to predict and prioritize new biomolecular annotations are both effective and efficient, especially if compared with the 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 here a new element that can improve the accuracy and relevance of the results of an annotation prediction and prioritization pipeline. We propose a *novelty indicator* able to state the level of "newness" (or "originality") of the annotations predicted for a specific gene to Gene Ontology terms, and to help prioritizing the most *novel* and *interesting* annotations predicted. We performed a thorough biological functional analysis of the prioritized annotations predicted with high accuracy by using this indicator and our previously proposed prediction algorithms. The relevance of our biological findings proves the effectiveness and trustworthiness of our proposed indicator and of its prioritization of annotation prediction pipeline results.

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

1 Introduction

In the past ten years, genomic data and information have incredibly grown [1] [2], creating a lot of new opportunities for scientists and biomedical researchers, especially in computational biology. To better express available biological knowledge and effectively use it to analyze these genomics big data, computational biologists use 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 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 hypothesis in biology. Many computational tools (e.g. GFINDer [4] [5], 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 may contain several errors, because only a small percentage of these annotations are supervised by human curators. Third, since different laboratoriess around the world may work on the same genes or proteins and reach different discoveries, annotation databases may contain inconsistent or ambiguous information about the same genes.

In this context, a key role is played by computational techniques, based on 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

- *D. Chicco is with the Princess Margaret Cancer Centre, University of Toronto, Ontario, Canada. <davide.chicco@davidechicco.it>*
- *F. Palluzzi is with the Dipartimento di Elettronica Informazione e Bioingegneria, Politecnico di Milano, and Istituto Europeo di Oncologia (IEO), Milan, Italy. <fernando.palluzzi@gmail.com>*
- *M. Masseroli is with the Dipartimento di Elettronica Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy. <masseroli@elet.polimi.it>*

computational intelligence algorithms for this goal [11]-[20]. 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, such as *cellular process*. This issue makes these 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 "newness" (or the "originality") of annotations predicted for a gene, and thus helping their prioritization. Furthermore, we present the performed thorough functional analysis of the biological relevance of the main prioritized prediction results obtained using the proposed indicator

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, and the techniques that we used to obtain and select them in Section 4.1. Then, biological relevance 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 context. semantic similarity between two genes or proteins. their GO annotations. In a survey by Pesquita et al. [21], the authors comprehensively described all the main semantic similarity measures used in the biomedical ontology domain, providing also some examples of implementations, applications, and a complete comparison. In particular, they described the Jiang [22], Lin [23] and Resnik [24] rates, that 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 analyzed 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 rates, 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-theroretic knowledge with other available biological information. Lord and colleagues [25] introduced a protein similarity approach that combines protein sequence similarity (computed through bioinformatics tools such as BLAST [26]) and semantic similarity based on protein GO annotations (computed through classical measures such as the Resnik one [24]). Their measure generates interesting results, but does not consider the tree position of GO terms, and is very bounded to the GO Molecular Function sub-ontology. Conversely, Speer et al. [27] 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 may lead to very different similarity results.

Starting from the two similarity measures by Lord and colleagues [25] and Speer et al. [27], Schlicker et al. [28] properly modified the Lord [25] and Speer [27] rates, and developed a new indicator, named $\text{GOscore}_{\text{BM}}$, 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 rate 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 [28].

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 described in [20], 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), all described in [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) [29], [30], 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 [31]. 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 [29]. 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.

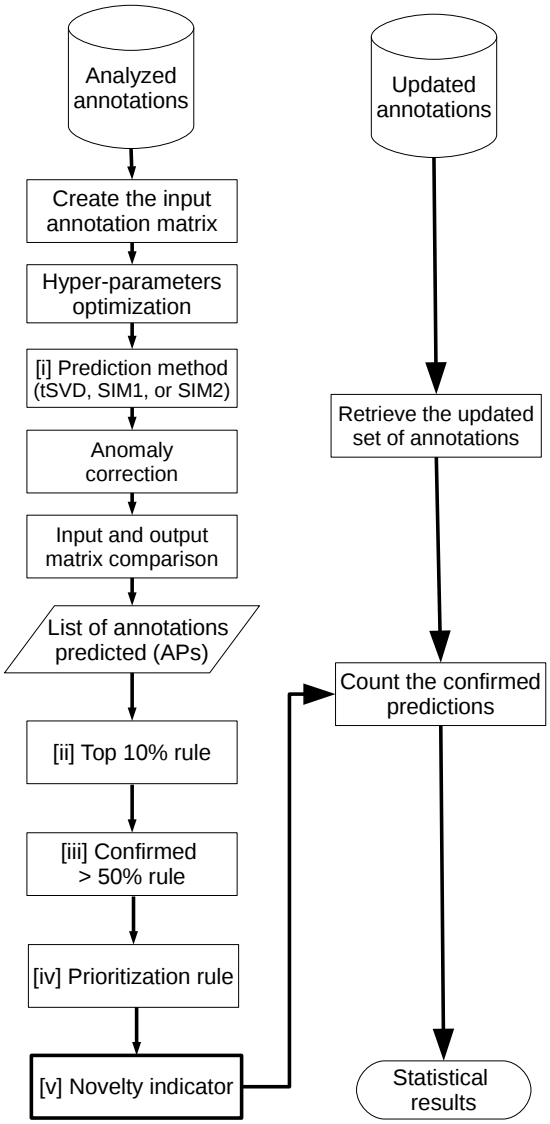

Fig. 1. Flowchart of the described computational pipeline for the prediction of biomolecular annotations (left hand side) and for their validation (right hand side). The [i] to [v] steps correspond to the controls listed in Section 4.1. The *novelty indicator* introduced in this paper is the [v] step on the left hand side of the flowchart.

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.

PRESENTE IN FIGURA

3.2 Novelty indicator

The $\text{GOscore}_{\text{BM}}$ semantic similarity measure was introduced to evaluate the similarity between two genes based on their GO annotations [28]. 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 $\text{GOscore}_{\text{BM}}$ is 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 this measure is to build a matrix S with the binary semantic similarity measures between each term of the first set GO^p and each term of 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 $i = 1, ..., N$ rows and $j = 1, ..., M$ columns, each matrix element s_{ij} is computed as the Resnik similarity measure (as defined in [24]) between the i_{th} element of the GO^p set and the j_{th} element of the GO^q set.

$$\forall i \in \{1, ..., N\},\ \forall j \in \{1, ..., M\} : \\ s_{ij} = ResnikSimilarity(\ GO_i^p,\ GO_j^q\ ) \quad (1)$$

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

$$\forall i \in \{1, ..., N\}, \forall j \in \{1, ..., M\} : \\ rowMaxScore_i = max(s_{ij}) \\ \forall i \in \{1, ..., N\} : rowMaximaSum = \\ \sum rowMaxScore_i \quad (2)$$

$$\forall j \in \{1, ..., M\}, \forall i \in \{1, ..., N\} : \\ columnMaxScore_j = max(s_{ij}) \\ \forall j \in \{1, ..., M\} : columnMaximaSum = \\ \sum columnMaxScore_j \quad (3)$$

$$rowScore\ =\ rowMaximaSum\ /\ N \\ columnScore\ =\ columnMaximaSum\ /\ M \quad (4)$$

The $GOscore_{BM}$ measure for the genes p and q is then defined as:

$$GOscore_{BM}(p, q) = \\ max(rowScore(p, q), columnScore(p, q)) \quad (5)$$

We use this score to measure the level of "newness" of the annotations predicted for a gene g, by comparing the set of GO terms associated with the gene g before the prediction (g_{before}) and the set of terms associated with it after the prediction (g_{after}). The more the two sets are different, the lower the $\text{GOscore}_{\text{BM}}$ is. Among all the most common semantic similarity measures available between two terms of an ontology (Jiang [22], Lin [23] and Resnik [24]), we decided to use the Resnik one because it is considered the most efficient rate in correlating gene sequence similarities [32] [33]. Since the Resnik similarity measure has no upper bound, the $\text{GOscore}_{\text{BM}}$ that uses it has no predefined upper bound; this does not influence our application, since we look for low values of the score, which we heuristically defined as $\text{GOscore}_{\text{BM}} < 1$.

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 explained in [19], 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 descriptive features, such as *cell cycle*. 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.

TABLE 1
Quantitative characteristics of the Homo sapiens GO annotations predicted in the tests. tSVD, SIM1 and SIM2 are prediction algorithms described in [12]. Cellular Component, Molecular Function and Biological Process are the sub-ontologies of the Gene Ontology.

method	Cellular Component	Molecular Function	Biological Process	total
tSVD	8	81	112	
SIM1	8	13	116	
SIM2	8	30	111	
total				600

3.3 Novelty indicator test

We tested the use of the $\text{GOscore}_{\text{BM}}$ measure as a novelty indicator on all the GO annotations predicted for the *Homo sapiens* genes, based on the gene GO annotations available in the GPDW dataset of July 2009 and using the (tSVD), Semantically IMproved tSVD with gene clustering (SIM1), Semantically IMproved tSVD with gene clustering and feature term similarity weights (SIM2) tSVD, SIM1 and SIM2 methods described in [12]. The $\text{GOscore}_{\text{BM}}$ measure resulted in concordance with the visual evaluation of the gene annotations performed by an expert. As a general example, we show the DAG of the GO Biological Process terms annotated to the human *protein phosphatase methylesterase 1* (*PPME1*) gene in Figure 2. The annotation prediction for this gene lead to a very low (good) zero indicator value ($\text{GOscore}_{\text{BM}}$ = 0.087).

4 RESULTS

We used the just-described *novelty indicator* as the last step of our computational pipeline for annotation prediction, prediction techniques, 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.

4.1 Prioritized predicted annotations

Fig. 2. Directed acyclic graph (DAG) of the GO Biological Process terms annotated to the *Homo sapiens PPME1* gene (Entrez Gene ID *51400*). The black circles represent the terms already known to be annotated to the human *PPME1* gene before the prediction, while the blue hexagons represent the terms of the predicted annotations using the tSVD method with our optimized parameters [17].

We ran some prediction tests on the Homo sapiens GO annotations of the previously-mentioned GPDW dataset of July 2009, by using the tSVD, SIM1 and SIM2 methods described in [12]. We report the quantitative characteristics of the predictions in Table 1.

On the basis of the prediction algorithms measures and *novelty indicator* described or referenced in the previous section, we can obtain a list of the most likely predicted annotations, which can be prioritized according to the accuracy conditions following described: usable by biologists and physicians to address their experiments, and here we report this directory for the *Homo sapiens* annotations. We report this list in Table 2 annotations, not found in the updated version of the database, that respect these conditions: we list such GO Biological Process predicted annotations for the considered 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) described in [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;
[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 con-

sidered 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 [20];

[v] Low *novelty indicator* value ("$GOscore_{BM} < 1$" column in Table 2), which indicates a very novel prediction for the gene (this condition is true for all the genes in Table 2; we heuristically evaluate *novel enough to be considered* only the predictions leading to novelty indicator values lower than 1.

We report all the just-mentioned prioritization conditions in Figure 1, where we label them with their corresponding numerals (from [i] to [v]).

In Table 2, we list the novel GO Biological Process annotations predicted for the *Homo sapiens* genes, based on the old GPDW version considered (of July 2009), and not found reported in the more updated GPDW database version considered (of March 2013); they satisfy both the *novelty indicator* condition and at least other three of the above prioritization conditions, and are sorted by decreasing number of conditions satisfied. Among all the 600 predicted GO annotations of Table 1, only 7 satisfy at least four items of the [i]-[v] accuracy conditions. Therefore, the GO annotations reported in Table 2 represent only the 1.17% of all the gene-function relationships predicted by our algorithms, and they would not be obtained as such without using the proposed *novelty indicator*. The *novelty indicator* condition states the relevance of the predictions: without it our pipeline would not be able to drop obvious and trivial gene functions, limiting its ability to lead to significant biological discoveries. The added value of the *novelty indicator* to our pipeline consists in constraining the predicted annotations to be "novel enough" to raise the interest of the biologists' community, as we address in functional analysis Section 4.2.

The predicted annotation list reported in Table 2 is the final biological relevance we can provide to physicians and biologists to address their experiments about human genes. A concrete application of our methods, that we hope may improve and quicken the discovery of new cures, new therapies, and new knowledge about gene functions. The annotations reported in Table 2 are 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 regards 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, and regards 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 SIM1 and SIM2 methods, but it regards a gene with more than half of the predicted annotations found confirmed in the more updated version considered of the GPDW database, and regards 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 regards 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 regards a quite general function and high 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 [34]-[36]. 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 already mentioned, 600 gene GO annotations were predicted, but only 7 of them (1.17%) satisfied

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. tools. The "*# conditions*" column states how many conditions of this table were satisfied by the specific gene prediction. 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 all the three 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 "$GOscore_{BM} < 1$" column states if the introduced *novelty indicator* has a low value for the annotated gene, that indicates a very novel prediction for the gene.

# conditions	Gene symbol (Entrez Gene ID)	GO term	predicted by			top 10%	conf. $> 50\%$	pred. conf.	$GOscore_{BM} < 1$
			tSVD	SIM1	SIM2				
6	PPME1 (51400)	Organelle organization. [BP] (GO:0006996)	√	√	√	√		√	√
6	CHST14 (113189)	Chondroitin sulfate proteoglycan biosynthetic process. [BP] (GO:0050650)	√	√	√		√	√	√
6	CHST14 (113189)	Biopolymer biosynthetic process. [BP] (GO:0043284)	√	√	√		√	√	√
5	CHST14 (113189)	Dermatan sulfate proteoglycan biosynthetic process. [BP] (GO:0050651)	√	√	√		√		√
5	CPA2 (1358)	Proteolysis involved in cellular protein catabolic process. [BP] (GO:0051603)		√	√		√	√	√
4	PPME1 (51400)	Chromosome organization. [BP] (GO:0051276)	√	√		√			√
4	CNOT2 (4848)	Positive regulation of cellular metabolic process. [BP] (GO:0031325)	√			√		√	√

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 described in the previous section prioritized the seven predicted GO Biological Process (BP) annotations for the four *homo sapiens* genes listed in Table 2. To assess the validity and biological relevance 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 [37]. 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 gene annotations, therefore improving current biological knowledge. As mentioned, the addition of the *novel indicator* presented in this paper allowed our computational prediction pipeline to avoid selecting obvious high–level descriptive features, and to finally prioritize annotations deemed "novel enough" to raise the biologists' community attention.

For each gene in Table 2, we retrieved a network of interacting neighbor genes from the STRING database [38]. We chose STRING for three reasons: (i) to take

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 predicted gene annotations prioritized. PFPs are predicted interactors of the annotated gene, according to STRING [38]. 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.

gene	Entrez Gene ID	GO BP	GO ID	biological validation	PFPs	ASGs	KEGG enrichment (p-value)
PPME1	51400	Organelle organization Chromosome organization	GO:0006996 GO:0051276	Primary evidence Secondary evidence	TJAP1 PPP2CA PPP2R1A PPP2CB PPP2R1B PPP4R1L PPP4C LCMT1 XRRA1 DNAJB13	AKT1 AXIN1 STRIP1 PPP2R4 PPP2R2A PPP2CA PPP2R1A PPP2CB TJAP1 LCMT1	hsa04530: Tight junction ($3.1E^{-7}$) hsa04310: Wnt signaling pathway ($9.3E^{-4}$) hsa04730: Long-term depression ($1.7E^{-3}$) hsa04350: TGF-beta signaling pathway ($2.5E^{-3}$) hsa04114: Oocyte meiosis ($4.0E^{-3}$)
CHST14	113189	Dermatan sulfate proteoglycan biosynthetic process Chondroitin sulfate proteoglycan biosynthetic process Biopolymer biosynthetic process	GO:0050651 GO:0050650 GO:0043284	Already confirmed Primary evidence No evidence	BCAN NCAN VCAN CSPG4 CSPG5 DCN B3GAT1 B3GAT2 ZNF469	BCAN NCAN VCAN CSPG4 CSPG5 DCN IGF1	hsa00532: Chondroitin sulfate biosynthesis ($1.6E^{-4}$)
CNOT2	4848	Positive regulation of cellular metabolic process	GO:0031325	Primary evidence	CNOT1 CNOT3 CNOT4 CNOT6 CNOT6L CNOT7 CNOT8 CNOT10 TOB1 RQCD1	CPEB3 PAIP1 CNOT1 CNOT8 TOB1 PABPC1 UBC UBE2D1 UBE2E1 PSMA1 PSMA2 PSMA3 PSMA4 PSMA6 PSMC3 PSMD7	hsa03018: RNA degradation ($1.3E^{-8}$) hsa03050: Proteasome ($2.2E^{-8}$)
CPA2	1358	Proteolysis involved in cellular protein catabolic process	GO:0051603	Secondary evidence	CTRB1	UBC CTRB2 POR LNX SLC9B1 SLC9B2 CA2 ATP6V0D1 ATP6V0D2	

advantage of the different data types integrated in STRING, including experimental assays, inferred interactions, and text mining; (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 the considered gene x and annotated to at least one 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 in Table 2.

For every considered gene, each annotation term 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:0050651)>, was available (i.e. confirmed) at the time of writing. The CHST14 gene annotation to *chondroitin sulfate proteoglycan biosynthetic process* (GO:0050650), 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 [39]-[42].

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, 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, XRRA1 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 PPME1 predicted annotation, to *chromosome organization* (GO:0051276), 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:0051276) is found in four distantly 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

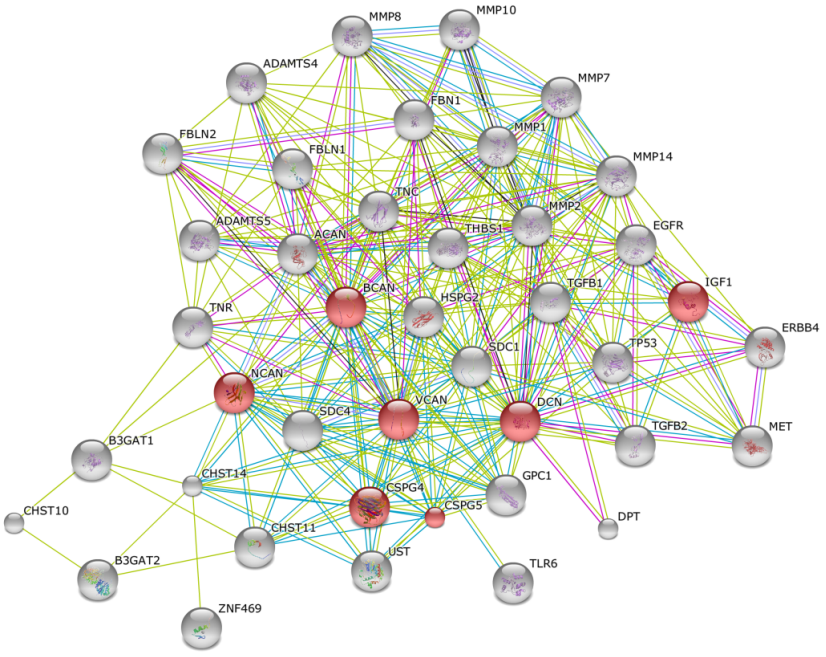

Fig. 3. Expanded (at depth 3) interaction network of the human CHST14 gene. Enrichment was found for the annotation to *chondroitin sulfate proteoglycan biosynthetic process* (GO:0050650). Dark red nodes support the predicted annotation. Color code of the connections is in accordance with the one used in STRING [38] (green: neighborhood; red: gene fusion; blue: co-occurrence; black: co-expression; purple: experiment-based; cyan: database-based; light green: text mining; violet: homology).

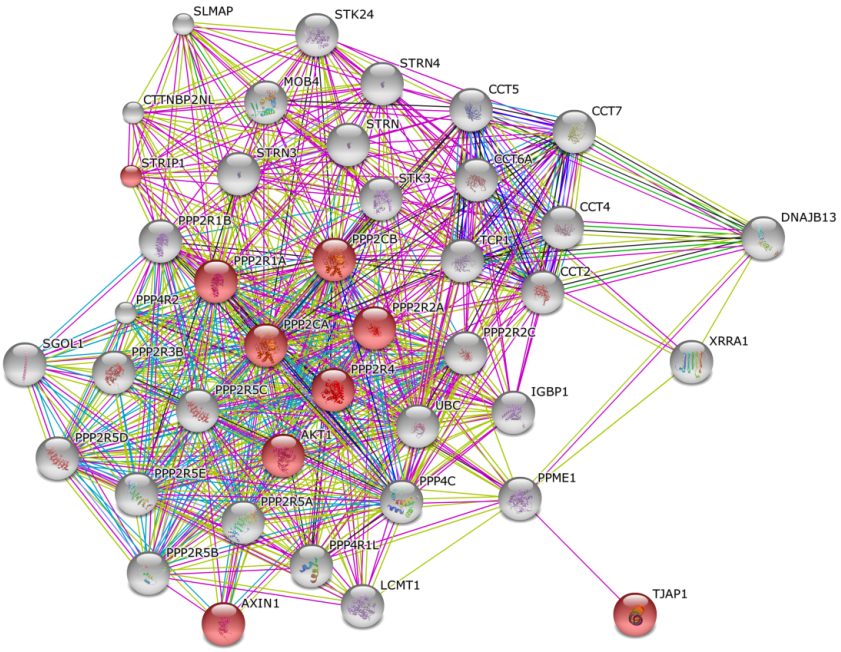

Fig. 4. Expanded (at depth 3) interaction network of the human PPME1 gene. Enrichment was found for the annotation to *organelle organization* (GO:0006996). Dark red nodes support the predicted annotation. Color code of the connections is as in STRING [38] (green: neighborhood; red: gene fusion; blue: co-occurrence; black: co-expression; purple: experiment-based; cyan: database-based; light green: text mining; violet: homology).

diseases including breast cancer [43] and Alzheimers disease [44]. 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 interac-

tion 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 encodes for a precursor of Ubuquitin and interacts with POR and ATP6V0D1, two ASGs of CPA2. CPA2 is a pancreatic carboxipeptidase involved in insulin metabolism [45]. In our prioritized prediction, CPA2 has been annotated with proteolitic 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 CPA2 proteolytic activity is documented since long time [46], [47], the relationship with insulin metabolism has not been clarified yet. In a recent work [48], it has been shown how high expression levels of genes (including CPA1, CPA2 and 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 CPA2 to its ASGs (i.e. insulin metablism 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 works by introducing a *novelty indicator* able to state the level of "newness" (or "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 a thorough 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. We also aim to integrate our computational pipeline into the online Bio Search Computing framework [49], [50], publicly available through the internet.

References

[1] D. J. Rigden, X. M. Fernndez-Surez, and M. Y. Galperin, "The 2016 database issue of Nucleic Acids Research and an update Molecular Biology Database Collection". *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1–D6, 2016.

[2] EMBL Nucleotide Archive statistics, http://www3.ebi.ac.uk/Services/DBStats/

[3] The Gene Ontology Consortium, "Creating the Gene Ontology resource: Design and implementation", *Genome Res.*, vol. 11, no. 8, pp. 1425-1433, 2001.

[4] M. Masseroli, D. Martucci, and F. Pinciroli, "GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining". *Nucleic Acids Res.*, vol. 32, no. Web Server issue, pp. W293–W300, 2004.

[5] M. Masseroli, "Management and analysis of genomic functional and phenotypic controlled annotations to support biomedical investigation and practice." *IEEE Trans. Inf. Technol. Biomed.*, vol. 11, no. 4, pp. 376–385, 2007.

[6] F. Al-Shahrour, P. Minguez, J. Tarraga, I. Medina, E. Alloza, D. Montaner, and J. Dopazo, "FatiGO+: A functional profiling tool for genomic data. integration of functional annotation, regulatory motifs and interaction data with microarray experiments". *Nucleic Acids Res.*, vol. 35, no. Web Server issue, pp. W91–W96, 2007.

[7] D. W. Huang, B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki, "DAVID Bioinformatics Resources: Expanded annotation database and novel algorithms to better extract biology from large gene lists". *Nucleic Acids Res.*, vol. 35, no. Web Server issue, pp. W169–W175, 2007.

[8] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan, R. Apweiler, "QuickGO: a web-based tool for Gene Ontology searching". *Bioinformatics*, vol. 25, no. 22, pp. 3045-3046, 2009.

[9] D. W. Huang, B. T. Sherman, and R.A. Lempicki, "Bioinformatics Enrichment tools: Paths toward the comprehensive functional analysis of large gene lists". *Nucleic Acids Res.*, vol. 37, no. 1, pp. 1–13, 2009.

[10] G. Pandey, V. Kumar, and M. Steinbach, "*Computational approaches for protein function prediction: A survey*". Twin Cities: Department of Computer Science and Engineering, University of Minnesota, 2006.

[11] M. Tagliasacchi, and M. Masseroli, "Anomaly-free prediction of Gene Ontology annotations using Bayesian networks". *Proceedings of BIBE 2009*, IEEE, pp. 107–114, 2009.

[12] P. Pinoli, D. Chicco, and M. Masseroli. "Computational algorithms to predict Gene Ontology annotations." *BMC Bioinformatics*, vol. 16, suppl. 6, S4, 2015.

[13] D. Chicco, M. Tagliasacchi, and M. Masseroli, "Biomolecular annotation prediction through information integration". *Proceedings of CIBB 2011*, Universitá di Salerno, pp. 1–8, 2011.

[14] M. Masseroli, D. Chicco, and P. Pinoli, "Probabilistic Latent Semantic Analysis for prediction of Gene Ontology annotations". *Proceedings of IJCNN 2012*, IEEE, pp. 2891–2898, 2012.

[15] P. Pinoli, D. Chicco, and M. Masseroli, "Enhanced probabilistic Latent Semantic Analysis with weighting schemes to predict genomic annotations". *Proceedings of BIBE 2013*, IEEE, pp. 1–4, 2013.

[16] P. Pinoli, D. Chicco, and M. Masseroli, "Improved biomolecular annotation prediction through weighting scheme methods". *Proceedings of CIBB 2013*, University of Salerno, pp. 1–12, 2013.

[17] D. Chicco, and M. Masseroli, "A discrete optimization approach for SVD best truncation choice based on ROC curves". *Proceedings of BIBE 2013*, IEEE, pp. 1–4, 2013.

[18] P. Pinoli, D. Chicco, and M. Masseroli, "Latent Dirichlet Allocation based on Gibbs Sampling for gene function prediction". *Proceedings of CIBCB 2014*, IEEE, pp. 1–8, 2014.

[19] D. Chicco, P. Sadowski, and P. Baldi. "Deep autoencoder neural networks for Gene Ontology annotation predictions". *Proceedings of BCBHI*, ACM, pp. 533-540, 2014.

[20] D. Chicco, and M. Masseroli, "Ontology-Based prediction and prioritization of gene function annotations". *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 13, no. 2, pp. 248–260, 2016.

[21] C. Pesquita, D. Faria, A. O. Falcao, P. Lord, and F. M. Couto. "Semantic similarity in biomedical ontologies", *PLoS Comput. Biol.*, vol. 5, no. 7, e1000443, 2009.

[22] J. J. Jiang, and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy". *Proceedings of ROCLING97*, ACLCLP, pp. 19–33, 1997.

[23] D. Lin, "An information-theoretic definition of similarity". *Proceedings of ICML 1998*, vol. 98, pp. 296–304, 1998.

[24] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy". *Proceedings of IJCAI'95*, vol. 1, pp. 448–453, 1995.

[25] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble, "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation." *Bioinformatics*, vol. 19, no. 10, pp. 1275-1283, 2003.

[26] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSIBLAST: a new generation of protein database search programs." *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.

[27] N. Speer, C. Spieth, and A. Zell, "A memetic clustering algorithm for the functional partition of genes based on the Gene Ontology." *Proceedings of CIBCB 2004*, IEEE, pp. 252–259, 2004.

[28] A. Schlicker, F. S. Domingues, J. Rahnenfhrer, and T. Lengauer, "A new measure for functional similarity of gene products based on Gene Ontology." *BMC Bioinformatics*, vol. 7, no. 302, pp. 1-16, 2006.

[29] M. Masseroli, A. Canakoglu, and S. Ceri, "Integration and querying of genomic and proteomic semantic annotations for biomedical knowledge extraction". *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 13, no. 2, pp. 209–219, 2016.

[30] F. Pessina, M. Masseroli, and A. Canakoglu, "Visual composition of complex queries on an integrative Genomic and Proteomic Data Warehouse". *Engineering*, vol. 5, no. 10B, pp. 94–98, 2013.

[31] D. Chicco, and M. Masseroli, "Validation procedures for predicted Gene Ontology annotations". *Proceeding of CIBB 2015*, University of Salerno, pp. 1–6, 2015.

[32] X. Guo, R. Liu, C. D. Shriver, H. Hu, and M. N. Liebman, "Assessing semantic similarity measures for the characterization of human regulatory pathways". *Bioinformatics*, vol. 22, no. 8, pp. 967–973, 2006.

[33] N. Diaz-Diaz, and J. S. Aguilar-Ruiz, "GO-based functional dissimilarity of gene sets". *BMC bioinformatics* vol. 12, no. 1, 360, 2011.

[34] K. Xu, Y. Bai, A. Zhang, Q. Zhang, and M. G. Bartlam, "Insights into the structure and architecture of the CCR4-NOT complex". *Front. Genet.*, vol. 5, 137, 2014.

[35] T. Inada, and S. Makino, "Novel roles of the multifunctional CCR4-NOT complex in post-transcriptional regulation". *Front. Genet.*, vol. 5, 135, 2014.

[36] A. Stroynowska-Czerwinska, A. Fiszer, and W. J. Krzyzosiak, "The panorama of miRNA-mediated mechanisms in mammalian cells". *Cell. Mol. Life Sci.*, vol. 71, no. 12, pp. 2253–70, 2014.

[37] M. Kaneisha, S. Goto, S. Sato, Y. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in KEGG.". *Nucleic Acid Res.*, vol. 42, no. Database issue, pp. D199–D205, 2014.

[38] D. Szclarczyc, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovich, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. von Mering, "STRING v10: protein-protein interaction networks, integrated over the tree of life". *Nucleic Acid Res.*, vol. 43, no. Database issue, pp. D447–D452, 2015.

[39] A. C. Petrey, and C. A. de la Motte, "Hyaluronan, a crucial regulator of inflammation". *Front. Immunol.*, vol. 5, 101, 2014.

[40] A. Malmström, B. Bartolini, M. A. Thelin, B. Pacheco, and M. J. Maccarana, "Iduronic acid in chondroitin/dermatan sulfate: biosynthesis and biological function". *Histochem. Cytochem.*, vol. 60, no. 12, pp. 916–925, 2012.

[41] E. Rönnberg, F. R. Melo, and G. J. Pejler, "Mast cell proteoglycans". *Histochem. Cytochem.*, vol. 60, no. 12, pp. 950–962, 2012.

[42] P. W. Noble, F. R. Lake, P. M. Henson, and D. W. Riches, "Hyaluronate activation of CD44 induces insulin-like growth factor-1 expression by a tumor necrosis factor-alpha-dependent mechanism in murine macrophages". *J. Clin. Invest.*, vol. 91, no. 6, pp. 2368–2377, 1993.

[43] S. Baldacchino, C. Saliba, V. Petroni, A. G. Fenech, N. Borg, and G. Grech, "Deregulation of the phosphatase, PP2A is a common event in breast cancer, predicting sensitivity to FTY720". *EPMA J.*, vol. 5, no. 1, 3, 2014.

[44] J. M. Sontag, and E. Sontag, "Protein phosphatase 2A dysfunction in Alzheimer's disease". *Front. Mol. Neurosci.*, vol. 7, 16, 2014.

[45] V. Pérez-Vázquez, J. M. Guzmán-Flores, D. Mares-Álvarez, M. Hernández-Ortiz, M. H. Macías-Cervantes, J. Ramírez-Emiliano, and S. Encarnación-Guevara, "Differential proteomic analysis of the pancreas of diabetic db/db mice reveals the proteins involved in the development of complications of diabetes mellitus". *Int. J. Mol. Sci.*, vol. 15, no. 6, pp. 9579–9593, 2014.

[46] E. S. Trombetta, and I. Mellman, "Cell biology of antigen processing in vitro and in vivo". *Annu. Rev. Immunol.*, vol. 23, pp. 975–1028, 2005.

[47] K. L. Rock, I. A. York, and A. L. Goldberg, "Post-proteasomal antigen processing for major histocompatibility complex class I presentation". *Nat. Immunol.*, vol. 5, no. 7, pp. 670–677, 2004.

[48] S. Jayaraman, A. Patel, A. Jayaraman, V. Patel, M. Holterman, and B. Prabhakar, "Transcriptome analysis of epigenetically modulated genome indicates signature genes in manifestation of type 1 diabetes and its prevention in NOD mice". *PLoS One*, vol. 8, no. 1, e55074, 2013.

[49] D. Chicco, "Integration of bioinformatics web services through the search computing technology". Dipartimento di Elettronica e Informazione, Politecnico di Milano, Technical Report, 2012.

[50] D. Chicco, and M. Masseroli, "Software suite for gene and protein annotation prediction and similarity search". *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 12, no. 4, pp. 837–843, 2015.

Davide Chicco obtained his Bachelor of Science and Master of Science degrees 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 fellow at the Princess Margaret Cancer Centre (University of Toronto, Ontario, Canada). His research topics focus upon mainly machine learning algorithms applied to bioinformatics.

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 Ph.D. student at the Dipartimento di Elettronica, 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.

Marco Masseroli received the Bachelor and Master of Science degree in electronic engineering in 1990 from Politecnico di Milano (Milan, Italy), and a PhD in biomedical engineering in 1996, from Universidad de Granada (Granada, Spain). He is associate professor and lecturer of bioinformatics and biomedical informatics at the Dipartimento di Elettronica, Informazione e Bioingegneria of Politecnico di Milano. His research interests are in the area of bioinformatics and biomedical informatics, focused on biomolecular databases, controlled biomedical terminologies and ontologies to effectively retrieve, manage, analyze, and semantically integrate genomic information with patient clinical and high-throughout genomic data. He is the author of more than 170 scientific articles, which have appeared in international journals, books and conference proceedings.