Methods 93 (2016) 15–23

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Enhancing protein function prediction with taxonomic constraints – The Argot2.5 web server

Enrico Lavezzo [a,1], Marco Falda [a,1], Paolo Fontana [b], Luca Bianco [b], Stefano Toppo [a,*]

[a] Department of Molecular Medicine, University of Padova, Padova, Italy
[b] Istituto Agrario San Michele all'Adige Research and Innovation Centre, Foundation Edmund Mach, Trento, Italy

ARTICLE INFO

Article history:
Received 27 April 2015
Received in revised form 14 August 2015
Accepted 25 August 2015
Available online 28 August 2015

Keywords:
Gene Ontology
Automated protein function prediction
Taxon constraints
Semantic similarity

ABSTRACT

Argot2.5 (Annotation Retrieval of Gene Ontology Terms) is a web server designed to predict protein function. It is an updated version of the previous Argot2 enriched with new features in order to enhance its usability and its overall performance. The algorithmic strategy exploits the grouping of Gene Ontology terms by means of semantic similarity to infer protein function. The tool has been challenged over two independent benchmarks and compared to Argot2, PANNZER, and a baseline method relying on BLAST, proving to obtain a better performance thanks to the contribution of some key interventions in critical steps of the working pipeline. The most effective changes regard: (a) the selection of the input data from sequence similarity searches performed against a clustered version of UniProt databank and a remodeling of the weights given to Pfam hits, (b) the application of taxonomic constraints to filter out annotations that cannot be applied to proteins belonging to the species under investigation. The taxonomic rules are derived from our in-house developed tool, FunTaxIS, that extends those provided by the Gene Ontology consortium. The web server is free for academic users and is available online at http://www.medcomp.medicina.unipd.it/Argot2-5/.

1. Introduction

The functional annotation of gene products is a crucial step for understanding the biology of living organisms, in all its physiological and pathological aspects. Since 2000, the Gene Ontology Consortium (GOC) has provided a powerful resource to collect the multitude of known functions in a structured vocabulary, the Gene Ontology [1], which is organized as a directed acyclic graph and facilitates the access to functional data through automatic tools.

The development of bioinformatics methods for gene products annotation is an active field of research and an international challenge is held periodically to assess such methods and provide a snapshot of the state of the art [2]. Automated tools predict function using different criteria, but most of them take advantage of sequence similarity based approaches to find matches between the input gene or protein and a database of already characterized entities (either at whole sequence or domain level), in order to transfer functional features [3,4].

Despite the considerable effort of the scientific community, the main outcome from the first Critical Assessment of Function Annotation (CAFA) is that there is significant room for improvement, because all automatic pipelines that participated in the challenge suffered from lack of both recall and precision [2]. This problem also affects annotations in the Gene Ontology Annotation database (GOA), in particular those generated automatically: even though they are a valuable resource for many proteins and many organisms [5], their error rate is difficult to be quantified and even controlled. The GOC is constantly trying to limit this phenomenon by implementing a number of automatic checks which verify both file formats and partially data coming from annotation submitters. For example, back in 2011 they introduced an "Annotation black list" which specifies protein:GO term combinations that are not allowed as annotations [6]. Furthermore, Deegan et al. [7] proposed a more general approach to prevent incorrect associations between certain functions and specific taxa, providing a list of "taxon constraints" which explicitly defines such incompatibilities. Nevertheless, novel methods are needed to improve annotation quality, either by correcting existing errors or preventing the generation of novel ones. Our group already contributed to this research field through the development of Argot2 [6], a tool for the automated prediction of protein function which placed in the top ten of the best performing algorithms at CAFA and CAFA2. However, the context where the algorithm works has dramatically changed since the time of its initial development, in particular as

* Corresponding author.
E-mail address: stefano.toppo@unipd.it (S. Toppo).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ymeth.2015.08.021
1046-2023/

regards the size of the databanks: UniProt [8], for example, has nearly quadrupled the number of entries, while the amount of annotations in GOA has increased by a factor of six. Such expansion negatively impacts the tool performance, by affecting both the execution time and the management of intermediate steps. In addition, new resources have become available to improve the predictive ability, such as the taxonomic constraints mentioned before.

In this paper, we present Argot2.5 (Annotation Retrieval of Gene Ontology Terms), a tool designed for high throughput annotation of large sequence data sets which improves upon its predecessor [9] thanks to the implementation of multiple novel features: (1) a clustered version of UniProt is used for BLAST searches in order to remove redundancy and speed up the searching time; (2) a novel semantic similarity measure is adopted and tested; (3) an extended set of taxonomic constraints is applied according to the in-house developed tool FunTaxIS (http://www.medcomp.medicina.unipd.it/funtaxis), that expands the list provided by GOC [7].

2. Materials and methods

Argot2.5 algorithm is based on the Argot2 approach [9]. Briefly, the algorithm starts from DNA/protein sequences and performs a BLAST [10] search against UniProt [8] database and a HMMER3 [11] search against Pfam [12]. The results retrieved from these steps are used to query the GOA databank: the collected GO terms are ranked according to both the significance of the hit they come from (provided by the e-value) and their occurrence in the results. The terms are then further grouped by means of semantic similarity, computed with Lin's formula [13], and only the representatives of high-scoring clusters are reported. The final score of each predicted GO term is based on a combined measure that is called Total Score (TS). For further extended details refer to Falda et al. [9]. To address known issues of Argot2 we have implemented the following enhancements.

2.1. Clustering of UniProt databank

UniProt databank [8] is periodically updated and only proteins carrying GO annotations are kept and further processed. Proteins are clustered at 90% sequence identity and 80% overlap with the longest sequence of the cluster using CD-HIT [14]. The UniProt release adopted for all the analyses shown in this paper dates back to December 2014 and contains 89,136,540 proteins, 57,336,105 of which are annotated with at least one GO term. The CD-HIT step produced 11,435,228 clusters. The representatives of the clusters, called seeds, are used as sequence space for BLAST searches, thus reducing the computational time and mitigating the over-representation effect of almost identical sequences that saturate the BLAST hits list. Furthermore, this clustering step is supported by the high homogeneity of GO annotations, which characterizes proteins with 90% sequence identity or above, as recently reported in [15]. Finally, to prevent the loss of information related to non-representative sequences, each seed inherits all GO annotations of its cluster members.

2.2. Argot2.5 new input

The steps leading to the generation of Argot2 input have been revisited. In addition to the introduction of BLAST search against the clustered version of UniProt, a different weight has been given to Pfam hits. The latter have been downsized when highly similar and numerous BLAST hits are found.

2.3. Semantic similarity

Lin's semantic similarity [13] is known to suffer from the so called "shallow annotation problem" [16,17]. This is due to the use of the Most Informative Common Ancestor (MICA) in the formula to calculate the semantic similarity between two GO terms. As a result, two GO terms that are close to the root of the GO graph can yield a very high semantic similarity and are not distinguishable from a high-scoring pair of GO terms that are close to the leaves of the graph. For this reason, we have evaluated simGIC [18], a groupwise approach that should alleviate this phenomenon [17,19]. The implementation is given by the following formula:

$$\mathrm{simGIC}(GO_A, GO_B) = \frac{\sum_{t \in \{prop(GO_A) \cap Prop(GO_B)\}} \mathrm{IC}(t)}{\sum_{t \in \{prop(GO_A) \cup Prop(GO_B)\}} \mathrm{IC}(t)}$$

where GO_A and GO_B are two GO terms and their propagation up to the root is given by $prop(GO_A)$ and $prop(GO_B)$. simGIC is defined as the sum of the Information Content (IC) of each term t in the intersection of $prop(GO_A)$ and $prop(GO_B)$ divided by the sum of the IC of each term t in their union. IC has been calculated for each GO term according to Resnik [20]:

$$\mathrm{IC}(t) = -\log[p(t)]$$

where $p(t)$ is the probability of usage of the term in the corpus, which in our case corresponds to the sum of the occurrences of the term and its descendants in GOA database.

2.4. Taxonomic filtering

Before being processed by Argot2.5 algorithm, GO annotations retrieved by BLAST and HMMER searches are filtered using the taxonomic constraints provided by GOC [7] and the set of rules generated by the Functional Taxonomy Information System (FunTaxIS submitted, http://www.medcomp.medicina.unipd.it/funtaxis), an in-house developed tool for the automatic generation of taxon constraints. FunTaxIS algorithm works as follows: firstly, the frequency of association between GO terms and taxa is calculated by linking the protein accession identifiers in GOA databank to the taxonomic assignment of each protein. Then, for each GO term/taxon pair, their relative probability of association is calculated and is used to determine whether such association is allowed or not. This approach is able to considerably increase the number of constraints with respect to those provided by GOC (data not shown), which instead are manually generated. Starting from the taxon identifier specified by the user in the web form, GO terms are filtered according to the taxonomic rules related to the input species: forbidden GO terms are either deleted or replaced with allowed parent GO terms within a fixed edge distance. A breadth first search algorithm is applied to identify the closest potential substitutes that are allowed for that particular species, while the deletion is obtained by setting the substitution distance to zero, thus forbidding all replacements.

2.5. Benchmark

The performance of Argot2.5 has been assessed on two different data sets, one released after the first CAFA challenge and another one corresponding to the whole yeast (*S. cerevisiae*, NCBI Taxonomy: 4932) proteome. The CAFA test set consists of 531 and 587 proteins evaluated on the Molecular Function (MF) and Biological Process (BP) ontologies, respectively, and only their GO terms associated with an experimental evidence code in GOA have been selected as the sets of true annotations (804 terms for MF and 1608 for BP). Two additional methods have been added to this comparison: the recently published PANNZER [21] and a naïve