

# Gene Function Prediction in Five Model Eukaryotes Exclusively Based on Gene Relative Location Through Machine Learning

Flavio Pazos Obregón (✉ [fpazos@iibce.edu.uy](mailto:fpazos@iibce.edu.uy))

Instituto de Investigaciones Biológicas Clemente Estable

**Diego Silvera**

Instituto de Investigaciones Biológicas Clemente Estable

**Pablo Soto**

Instituto de Investigaciones Biológicas Clemente Estable

**Patricio Yankilevich**

Instituto de Investigación en Biomedicina de Buenos Aires (IBioBA), CONICET - Partner Institute of the Max Planck Society

**Gustavo Guerberoff**

Instituto de Matemática y Estadística "Prof. Ing. Rafael Laguardia", Facultad de Ingeniería, UDELAR

**Rafael Cantera**

Instituto de Investigaciones Biológicas Clemente Estable

---

## Research Article

**Keywords:** Gene function prediction, gene, machine learning, data sources, gene function, eukaryotes, organisms, Escherichia coli, E. coli

**Posted Date:** January 6th, 2022

**DOI:** <https://doi.org/10.21203/rs.3.rs-1199974/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

# Gene function prediction in five model eukaryotes exclusively based on gene relative location through machine learning

Flavio Pazos Obregón <sup>a, b</sup> (+) \*, Diego Silvera <sup>a</sup> (+), Pablo Soto <sup>a</sup>, Patricio Yankilevich <sup>c</sup>, Gustavo Guerberoff <sup>d</sup>, Rafael Cantera <sup>a</sup>

a – Departamento de Biología del Neurodesarrollo, Instituto de Investigaciones Biológicas Clemente Estable, Montevideo, Uruguay.

b– Unidad de Bioquímica y Proteómica Analíticas, Instituto Pasteur de Montevideo, Montevideo, Uruguay.

c – Instituto de Investigación en Biomedicina de Buenos Aires (IBioBA), CONICET – Partner Institute of the Max Planck Society, Buenos Aires, Argentina.

d – Instituto de Matemática y Estadística “Prof. Ing. Rafael Laguardia” , Facultad de Ingeniería, UDELAR, Montevideo, Uruguay.

(+) These authors contributed equally to this work

\* fpazos@iibce.edu.uy Av. Italia 3318, 11600 Montevideo, Uruguay

## Abstract

The function of most genes is unknown. The best results in automated function prediction are obtained with machine learning-based methods that combine multiple data sources, typically sequence derived features, protein structure and interaction data. Even though there is ample evidence showing that a gene's function is not independent of its location, the few available examples of gene function prediction based on gene location rely on sequence identity between genes of different organisms and are thus subjected to the limitations of the relationship between sequence and function.

Here we predict thousands of gene functions in five model eukaryotes (*Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus* and *Homo sapiens*) using machine learning models exclusively trained with features derived from the location of genes in the genomes to which they belong. Our aim was not to obtain the best performing method to automated function prediction but to explore the extent to which a gene's location can predict its function in eukaryotes. We found that our models outperform BLAST when predicting terms from Biological Process and Cellular Component Ontologies, showing that, at least in some cases, gene location alone can be more useful than sequence to infer gene function.

**Contact:** fpazos@iibce.edu.uy

**Supplementary information:** <http://gfpm1.bnd.edu.uy>

## 1. INTRODUCTION

We witness a growing gap between the number of assembled genomes and the number of genes with known functions. Less than 1% of the protein sequences in UniProtKB <sup>1</sup> have an experimental Gene Ontology annotation <sup>2</sup> and even in well studied organisms, the majority of known genes have yet no assigned function <sup>3</sup>. Furthermore, well studied genes have frequently been assigned more than one function so less studied genes, for which only one function is known, have probably more functions to be discovered <sup>4</sup>. In this context there is an increasing need to improve automated function prediction (AFP) <sup>5-9</sup>.

The Critical Assessment of protein Function Annotation algorithms (CAFA) is a series of experiments designed to provide a large-scale assessment of computational methods dedicated to automated function prediction (AFP) <sup>7,10,11</sup>. In all CAFA editions so far, the best results were obtained with machine learning-based methods combining multiple data sources, typically including sequence derived features, protein structure and molecular interaction data. The performance of the methods evaluated by the CAFA challenges improved dramatically between the first (2013) and the second (2016) edition but this improvement slowed down between the second and the third edition (2019). The authors hypothesized that including more varied sources of data will lead to additional large improvements in AFP <sup>7</sup>.

Thus, finding new ways to extract relevant biological information from the available data is key to improve AFP. For around 99% of all known proteins, the only available information is the sequence encoded in the corresponding genome, highlighting the importance of sequence-based AFP <sup>12</sup>. But AFP based on sequence similarity is hindered by a highly variable correlation between sequence identity and gene function <sup>13</sup> and by the evolutionary distance of many genomes to the closest well-characterized genome <sup>14</sup>. Here we explore the hypothesis that the relative location of a gene, a feature that is independent of sequence and can be directly extracted from any genome, is sufficient to perform AFP, at least with comparable performance to that reached by sequence similarity alone.

Functionally related genes may be constrained to remain close to each other due to natural selection, forming conserved gene clusters<sup>15</sup>. Local clusters of co-expressed, co-regulated or functionally related genes have been documented in a wide range of organisms, including prokaryotes, yeast, insects, vertebrates and plants<sup>16-23</sup>.

Equating conserved co-locality with co-functionality have been a fruitful approach for predicting gene functions in prokaryotes for more than 20 years<sup>15,24-28</sup>. On the contrary, there are very few examples<sup>14,29</sup> of the use of this approach in eukaryotic organisms, although also in eukaryotic genomes functions are non-randomly distributed<sup>21</sup>. However, these AFP studies were based on conserved gene neighborhoods, thus subjected to the limitations mentioned above regarding the relationship between sequence and function.

Here we performed AFP based exclusively on the relative location of genes. We test the predictive power of a feature which represents the spatial organization of genes with respect to their functions, which we term "functional landscape arrays" (FLAs). A FLA is an array associated to each gene that contains the enrichment in a set of Gene Ontology terms (GO terms) found around the gene, considering different window sizes. These arrays contain information which is independent of sequence similarity between genes and that can be automatically extracted from any annotated genome.

We predicted associations between genes of five eukaryotes (*Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus* and *Homo sapiens*) and terms from the three ontologies of the Gene Ontology (Biological Process, Cellular Component and Molecular Function) training a set of hierarchical multi-label classifiers with FLAs. Then we compared the results of our 15 models, one for each pair organism/ontology, with equivalent models that randomly assign functions to genes. We found that our models, trained exclusively with location-derived features, performed better than chance in the five organisms and in the tree ontologies, showing there is useful information in the way in which genes are distributed along these genomes.

We also compared the performance of our models to the performance of BLAST, one of the baseline methods of CAFA 3<sup>7</sup>. Our models outperform BLAST when predicting Biological Process in the three organisms for which specific data from the last CAFA is available. These results indicate that gene location can be as informative as sequence homology when performing AFP. These results also support the idea that gene distribution patterns are tightly regulated in eukaryotic genomes and its incorporation as predictive features could improve AFP.

## 2. METHODS

### 2.1 General procedure to predict associations between genes and GO terms

For each genome,

- Model as a string of protein coding genes.
- Random split in sets T and E, containing 80% and 20% of the genes

respectively.

For each Ontology,

- Train a binary classifier for each GO term X associated with at least 40 genes in T and 10 genes in E
  - Training set: genes in T annotated with GO term X (as positives) and its siblings (as negatives)
  - Predictive feature: a FLA for each gene, including enrichment in GO term X, its siblings and its ancestors
  - Hyper-parameters set by grid search & cross validation
- Combine all the binary classifications into one hierarchical multi-label classifier using the node interaction method.
- Evaluate calculating the hF1 score over the test set E
- Using the classification threshold that maximizes hF1 over E, predict new associations between GO terms and genes in E.

## 2.2 Genome modeling

We model the genome as a collection of segments (the chromosomal arms) in which the protein coding genes –the only elements we considered– are located one next to the other, without intergenic regions or superpositions <sup>30</sup>. In this model, the position of a gene is defined by the location of its transcription starting point and the distance between two genes is the number of other genes located between them. The number of protein-coding genes considered in each genome is shown in Table 1.

## 2.3 Gene Ontology

Gene Ontology (GO) is an attempt to describe all the knowledge about the biological functions of genes with three ontologies: Molecular Function, Cellular Component and Biological Process, each one representing different aspects of the biology of a gene product and organized as a directed acyclic graph <sup>2</sup>. Each “GO term” is a node of these graphs, with precise definition and relationships with other terms. A GO annotation occurs when an association between a gene product and a GO term is established. We used a version of the ontology downloaded on November 2018. To fulfill the true path rule <sup>31</sup>, given the annotations of an organism within a given ontology, we up-propagated all the annotations, meaning that if a gene was annotated with a given GO term we associated that gene with all the ancestor terms up to the root of the graph.

## 2.4 Local enrichment analysis

Enrichment analysis is a method frequently used to determine if a given gene feature is overrepresented in a list of genes <sup>32</sup>. It assesses if the genes of a list associated with a given feature are more frequent than what should be expected in a list of genes of the same size but randomly picked from the same background list.

Given a gene of interest  $j$ , we define the Local Enrichment in the GO term  $x$  for the gene  $j$  and a window  $w$  centered in  $j$  as:  $E_{jxw} = ((k/n) / (M/N))$ , where  $N$  is the number of genes in the chromosomal arm,  $M$  is the number of genes in the

chromosomal arm associated with GO term  $x$ ,  $n$  is the number of genes in the window and  $k$  is the number of genes in the window associated with GO term  $x$  (see Figure 1). In other words,  $E_{j,x,w}$  assess if the genes annotated with the GO term  $x$  are located in the surroundings of gene  $j$  more frequently than what could be expected by chance. This approach was successfully used to look for clusters of GO terms along the genome of seven eukaryotes <sup>33</sup>.

## 2.5 Functional Landscape Arrays and Functional Enrichment Maps

To functionally characterize the surrounding of a gene we calculated its local enrichment in various GO terms. We considered a window  $w$ , centered in the gene under consideration, that includes 5, 10, 20, 50 or 100 genes to each side of the gene. The window was moved stepwise one gene at a time until the entire chromosome was covered (see Figure 1). Then, for each gene we defined a Functional Landscape Array (FLA): an array with a row for each window size and a column for each GO term whose enrichment was evaluated. Because of computational limitations, in the work we are reporting here, the GO terms included in each FLA depend on the GO term to be classified: we only included the enrichment found in that GO term, its father, its siblings and all its descendants.

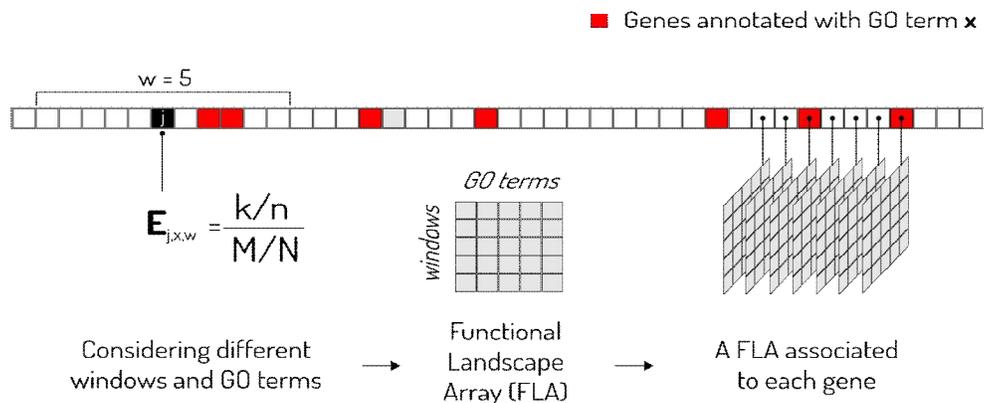


Figure 1. Local enrichment analysis and Functional Landscape Arrays.  $k$  is the number of genes in the window associated with GO term  $x$ ,  $n$  is the number of genes in the window,  $M$  is the number of genes (squares) in the chromosomal arm (strip) associated with GO term  $x$ , and  $N$  is the total number of genes in the chromosomal arm.

Importantly: to train our models we did not consider the annotations of the genes in the set E, that was reserved for the evaluation of the models. This procedure guarantees an unbiased evaluation of the classifiers, in which the features used in training are not extracted from examples used in testing. Nevertheless, because it is a useful result by itself, we also performed Local Enrichment Analysis along each genome considering all its current annotations. We calculated the local enrichment around all the genes in each genome using the same set of window sizes and for all those GO terms associated with at least 20 genes and obtained what we call "functional enrichment maps". The functional enrichment map of a given GO term shows which regions of a genome are enriched in that GO term, for various windows sizes.

## 2.6 Implementation of hierarchical multi label classifiers

We implemented a hierarchical multi label classifier for each pair organism / ontology using, with some modifications, the algorithm proposed in <sup>34,35</sup>. This is a local approach, since a binary classifier is trained for each GO term. Due to computational limitations, for the binary classification at each node, instead of a Support Vector Machine, we used a Random Forest classifier <sup>36</sup>, that have comparable performance in gene function prediction but with lower computational cost. For the same reason we did not used SMOTE <sup>37</sup>, a technique used to artificially generate new labeled data when training sets are too small. Depth, number of trees and measure of impurity for each classifier were set by grid search and 3-fold cross validation.

First we randomly split the genome into two sets: T and E. Set T includes 80% of the genes and was used to define the training sets and to obtain the FLAs. Set E includes the remaining 20% of the genes and was used to evaluate the models. We trained a binary classifier for each GO term that was associated with at least 40 genes in T and at least 10 genes in E. Table 2 shows the amount of GO terms meeting these conditions in each organism and ontology, i.e. the GO terms that could be predicted.

To define the training set for each classifier we applied the siblings policy <sup>38</sup>. We included as positive cases those genes associated with the GO term under consideration and as negative cases those genes associated with the siblings or uncles terms of the GO term under consideration and not to that term. Importantly, to construct the FLA associated to each gene, to be used as predictive feature, we only considered the annotations of the genes that belonged to **T**.

With each trained classifier we classified the genes in **E** and then post-processed the predictions using the node interaction method <sup>35</sup>, to respect the restrictions imposed by the hierarchy of the ontology. Finally, we evaluated the performance of each hierarchical multi-label classifier using the hierarchical version of the F1 score. All calculations were carried out using ClusterUY (site: <https://cluster.uy>).

## 2.7 Evaluation of the models

To evaluate the performance of each trained model we used the complete set of annotations of the genes in **E**, that were not used in training. As evaluation metric we used the hierarchical version of the F1 score (hF1) proposed in <sup>39</sup> and used in the CAFA competitions. If we denote the true and false positives as TP and FN and the true and false negatives as TN and FN, Precision (Pre) and Recall (Rec) are defined as:

$$Pre = \frac{TP}{TP+FN} \quad Rec = \frac{TP}{TP+FN}$$

and their hierarchical versions, which we term hPre and hRec, are defined as:

$$hPre(\theta) = \frac{\sum_{i=1}^n |Pi(\theta) \cap Ti|}{\sum_{i=1}^n |Pi(\theta)|} \quad hRec(\theta) = \frac{\sum_{i=1}^n |Pi(\theta) \cap Ti|}{\sum_{i=1}^n |Ti|}$$

where  $\theta \in [0,1]$  is the classification threshold,  $n$  is the number of genes,  $T_i$  is the set of GO terms truly associated to gene  $i$  and  $P_i(\theta)$  is the set of GO terms predicted for gene  $i$  with the classification threshold set at  $\theta$ . We

assumed that the root of each ontology always is in  $P_i(\theta)$ . The hF1 score is the harmonic mean of hPre and the hRec and is defined as:

$$hF1(\theta) = \frac{2 \cdot hPre(\theta) \cdot hRec(\theta)}{hPre(\theta) + hRec(\theta)}$$

## 2.8 Comparison with random models

As a way to assess how far from randomness is the distribution of gene functions along the genome, we compared the hF1 of each of our trained models with the hF1 reached by an equivalent model that randomly assigns GO terms to genes. In these “random models”, the probabilities of association between each gene of the corresponding genome and each GO term for which a binary classifier was trained are randomly selected from a uniform distribution. For each organism and ontology, we obtained the ratio between the hF1 of the trained model and the hF1 of its random version.

## 2.9 Comparison to one of the CAFA baseline methods

We also compared the performance of our models to the performance of BLAST, one of the baseline methods used in CAFA 3. In this case, BLAST was based on search results using the Basic Local Alignment Search Tool software against the training database<sup>40</sup>. A term was predicted as the highest local alignment sequence identity among all BLAST hits annotated with the term. BLAST was evaluated during CAFA 3 using the new experimental annotations accumulated during the competition (from February 2017 to November 2017). We used the same approach to compare the performance of these two CAFA methods to our method and used the new experimental annotations accumulated from November 2018 (date of the GO annotations that we had used to train our models) to September 2021.

We compared our models with BLAST when predicting GO terms for individual species. This data is available as Supplementary files for CAFA 3 at: <https://doi.org/10.6084/m9.figshare.8135393.v3> and includes performance evaluation for *H. Sapiens*, *M. musculus* and *D. melanogaster*. We compared our results with those obtained with the limited-knowledge benchmarks and under the

full evaluation mode. For more details about the different CAFA evaluations modes please refer to CAFA 3, Additional file 1 <sup>7</sup> and CAFA2 <sup>11</sup>

### 3. RESULTS

#### 3.1 Functional enrichment maps in 5 model eukaryotes

We performed Local Enrichment Analysis around each gene of a given genome considering windows of various sizes (See Methods). Local Enrichment Analysis of a given gene assess if the genes in the surroundings are annotated with any GO term more frequently than what could be expected by chance. Given a GO term, its functional enrichment map shows which regions of a genome are enriched in that GO term, considering various windows sizes. We obtained the functional enrichment map of all those GO terms associated with at least 20 genes in each of the five considered organisms. As an example, Figure 2 shows the functional enrichment maps in *D. melanogaster* of three GO terms that belong to the same branch of the Cellular Component ontology. The data to generate all the functional enrichment maps is available at: <https://github.com/IIBCE-BND/gfpml-datasets/tree/master/lea>

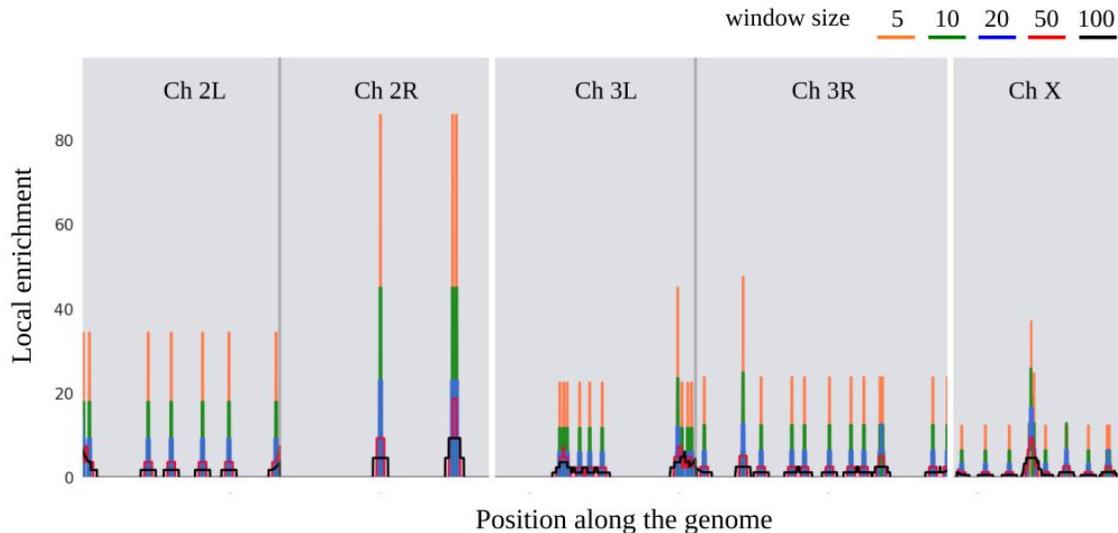


Figure 2. Functional enrichment map of the GO term “Golgi membrane” (GO:0000139) in the genome of *D. melanogaster*. There are 50 *Drosophila* genes annotated with this GO term that belongs to the Cellular Component ontology. The chromosomal position is represented in the x axis and the corresponding local enrichment at each position is shown in the y axis. Each light gray block corresponds to a chromosome (only chromosomes 2, 3 and X are shown) and the vertical dark gray lines mark the position of the centromeres, which divide the chromosome 2 into arms 2L and 2R and chromosome 3 into arms 3L and 3R. The enrichment found using different windows is shown with the colors indicated in the figure.

### 3.2 Implementation of hierarchical multi label classifiers

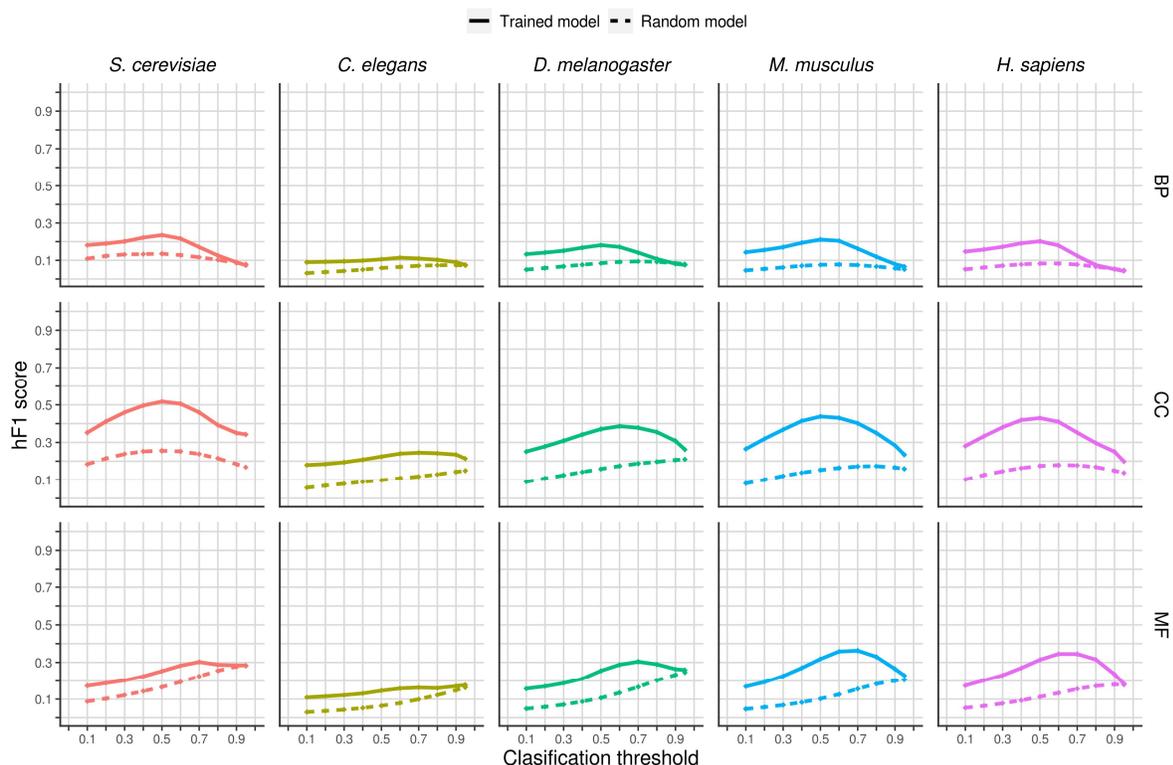
We trained fifteen hierarchical multi label classifiers, one for each possible pair organism/ontology. As detailed in Methods, we randomly split each genome into two sets: **T**, that includes 80% of the genes and was used for training, and **E**, that includes the remaining 20% of the genes and was used for evaluation. Each model assigned probabilities of association between the genes of the set **E** and those GO terms associated with at least 40 genes of the set **T** and 10 genes of the set **E**. Table 1 shows, for each organism and each ontology, the number of GO terms fulfilling these conditions and for which we implemented a binary classifier.

Organism	Proetin coding genes	Ontology	Total GO terms	Considered GO terms	hPrec	hRec	hF-max
<i>S. cerevisiae</i> (R64)	5,892	BP	5,074	525	0.24	0.23	0.24
		CC	1,035	137	0.51	0.52	0.52
		MF	2,323	137	0.69	0.19	0.30
<i>C. elegans</i> (WBcel235)	7,356	BP	5,661	551	0.09	0.15	0.11
		CC	1,110	117	0.19	0.33	0.25
		MF	2,226	151	0.25	0.14	0.17
<i>D. melanogaster</i> (BDGP6)	11,122	BP	7,416	880	0.17	0.20	0.18
		CC	1,277	176	0.41	0.37	0.39
		MF	2,599	212	0.47	0.22	0.30
<i>M. musculus</i> (GRCm38.p6)	20,809	BP	15,318	1040	0.22	0.21	0.21
		CC	1,953	285	0.46	0.42	0.44
		MF	4,269	364	0.63	0.25	0.36
<i>H. sapiens</i> (GRCh38.p13)	17,276	BP	13,816	1212	0.21	0.20	0.20
		CC	1,818	338	0.44	0.42	0.43
		MF	4,244	369	0.47	0.27	0.35

**Table 1.** GO terms for which a binary classifier was trained and tested. The first column shows the assembly version used for each organism, the second column shows the number of protein coding genes in each genome, the third column indicates the ontology, the fourth column shows the number of GO terms associated with at least one gene for that organism and ontology and the fifth column shows the number of GO terms associated with at least 40 genes in the set T (used for training) and 10 genes in the set E (used for evaluation). These are the GO terms for which a binary classifier was trained and tested. For each organism and ontology, we implemented a hierarchical multilabel classifier combining these binary classifiers. The hierarchical precision, recall and F-max reached by each of these models are shown in columns sixth, seventh and eight.

### 3.3 Evaluation of the models

We evaluated the performance of our models using the hierarchical version of the F1 score (hF1). Figure 3 shows the hF1 reached by each trained model over the test set E, as well as the hF1 of the corresponding random model, as a function of the classification threshold.



**Figure 3.** Hierarchical F1 over the test set for each model as a function of the classification threshold. In each plot the classification threshold, ranging from 0 to 1, is depicted in the x axis and the hF1, also ranging from 0 to 1, is depicted in the y axis. Trained models are represented by solid lines and random models by dotted lines. Each column of the panel corresponds to an organism and each row to an ontology (BP: Biological Process, CC: Cellular Component, MF: Molecular Function).

The hF-max is the highest hF1 score that the model reaches when varying the classification threshold. hF-max is a measure of the overall performance of the model and the corresponding classification threshold was used to predict new

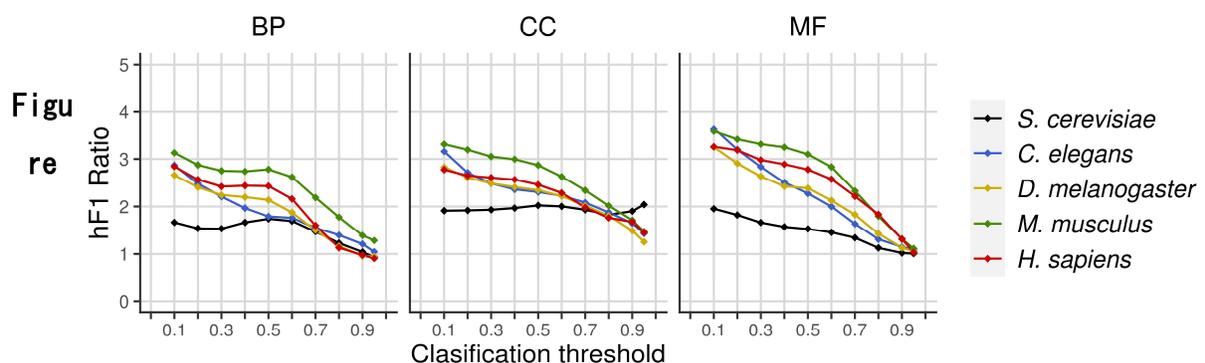
associations between GO terms and genes. Table 1 shows the hF-max for each model along with the corresponding precision and recall.

### 3.4 Comparison with random models

To assess how far from randomness is the linear organization of the genes along the genome with respect to its functions we calculated the ratio between the hF-max of the trained model and the hF-max of an equivalent random model, i.e. a model that randomly assigns probabilities of association between the same set of GO terms and the same genes (see Methods). Figures 4 and 5 show how this ratio varies with the classification threshold in each organism and ontology. Table 2 shows the hF-max reached by each model over the test set E. The trained models consistently performed better than the random models.

Organism	BP	CC	MF	Mean
<i>S. cerevisiae</i>	1.74	2.02	1.34	1.70
<i>C. elegans</i>	1.76	2.08	1.15	1.66
<i>D. melanogaster</i>	2.14	2.22	1.83	2.06
<i>M. musculus</i>	2.78	2.87	2.33	2.66
<i>H. sapiens</i>	2.43	2.46	2.57	2.49
Mean	2.17	2.33	1.84	

**Table 2.** Ratio between the hF-max reached by the trained model and the hF-max reached by the corresponding random model over the set E for each possible pair organism/ontology.



4. Ratio between the hF1 score of the trained model and the hF1 score of the corresponding random model as function of the classification threshold. Each graph shows the results for a given ontology, representing each organism with a different color.

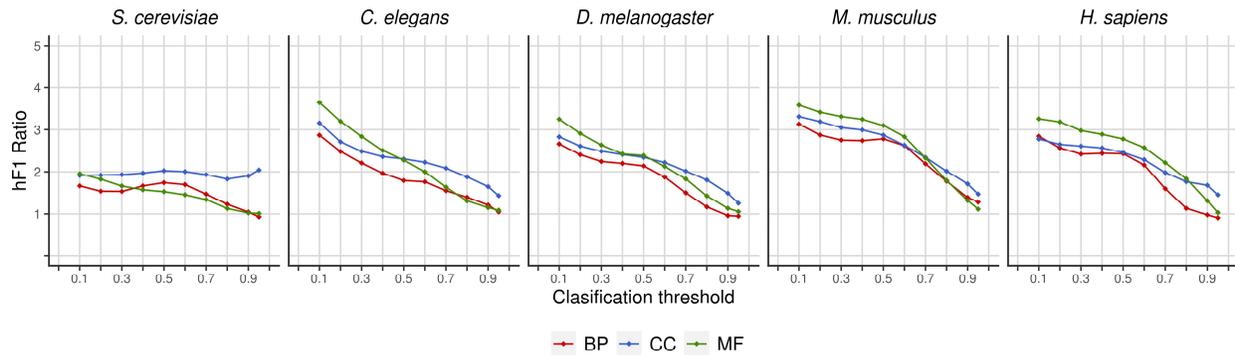


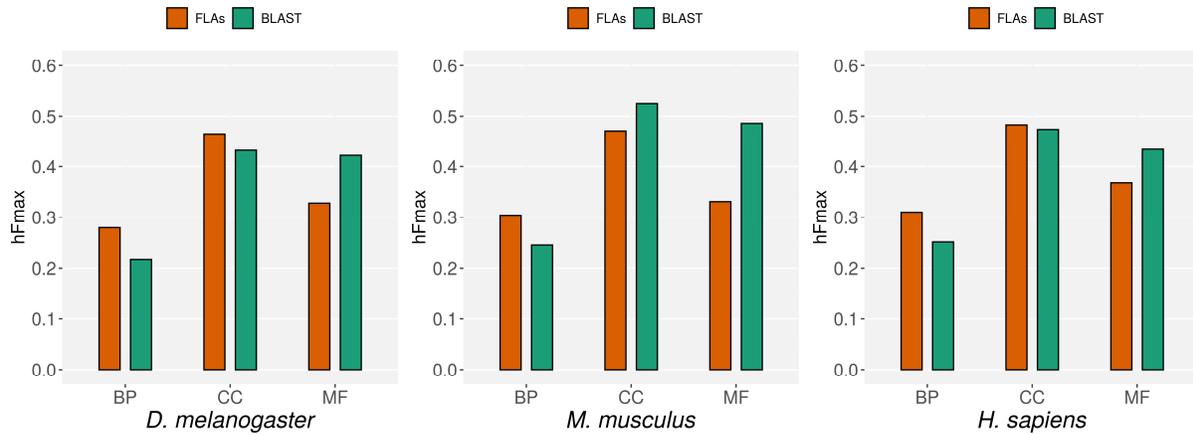
Figure 5. Ratio between the hF1 score of the trained model and the hF1 score of the corresponding random model as function of the classification threshold. Each graph shows the results for a given organism, representing each Ontology with a different color.

### 3.5 Comparison to one of the CAFA baseline methods

As a complementary way to evaluate our models we also compared their performance with the performance reached by BLAST, one of the baseline methods used in CAFA 3 (see Methods). Comparisons were made between the hFmax reached by our models and the hFmax reached by BLAST when making predictions for individual species, data that is only available for three of the five species we studied here: *H. sapiens*, *M. musculus* and *D. melanogaster*.

With this comparison we aim to assess if gene location alone can predict gene function with comparable performance to that reached by sequence homology alone. We found that this is the case and figure 6 shows the hFmax reached by the three models for each organism and ontology. Notably, for the three considered organisms the models trained with FLAs outperforms BLAST when predicting GO terms from the Biological Process Ontology. Our models also

outperform BLAST when predicting GO terms from the Cellular Component Ontology in *H. sapiens* and *D. melanogaster*.



**Figure 6.** Comparison to one of the CAFA baseline methods. Each graph shows the hFmax of different models when predicting GO terms of the three ontologies in three organisms. In red, the hFmax of the models exclusively trained with FLAs, evaluated using the new experimental annotations accumulated from November 2018 to September 2021. In blue, the hFmax of BLAST when making predictions on the same organisms and ontology as reported in CAFA 3<sup>7</sup>.

### 3.6 Prediction of new associations between genes and GO terms

We classified the genes in set **E** using the trained models and the classification threshold that maximizes the hF1 score (see Methods). We obtained the probability of association between each gene and each GO term associated with at least 40 genes in **T** and 10 genes in **E**. The complete set of predicted associations with a probability above the threshold is available at: <http://gfpml.bnd.edu.uy>. In this site, the user can browse and download all the predictions, searching by organism, ontology, chromosomal position, gene or GO term. Figure 7 shows a screen-shot of the site.

Choose organism:			
Homo sapiens			
Choose ontology:			
Biological process			
F-max of the mode: 0.35			
Gene ID	Chromosome	Starting position	GO term
			GO:0012501
ENSG00000261456	10p	0	GO:0012501
ENSG0000067064	10p	7	GO:0012501
ENSG0000067082	10p	12	GO:0012501
ENSG00000178363	10p	23	GO:0012501
ENSG00000134470	10p	29	GO:0012501
ENSG00000148429	10p	42	GO:0012501
ENSG00000165609	10p	48	GO:0012501
ENSG00000187522	10p	64	GO:0012501
ENSG00000077943	10p	73	GO:0012501

**Figure 7.** Screenshot of <http://gfpl.bnd.edu.uy>. At this website all the associations between genes and GO terms predicted by our models are available.

Once an organism and an ontology is chosen, the hFmax reached by the corresponding model as well as the list of predicted associations are shown.

#### 4. DISCUSSION

For the majority of the known genes, the only available information is their DNA sequence <sup>12</sup>. AFP based on DNA sequence similarity is a common approach, since its known that two genes with very similar sequences probably share function. But the contrary is not always true. A thorough study of the correlation between similarity in protein sequence and function in yeast <sup>13</sup> found that the majority of the sequences of proteins annotated with the same GO term were non-similar. In general, within one branch of an ontology tree, the more specific a GO term is, the more similar the sequences of the genes annotated with that term are, but the degree of similarity is highly variable and is significant only for specific GO terms. When using orthology between genes,

these methods face another limitation: the evolutionary distance of many genomes to the closest well-characterized genome. For example, only 25-50% of the proteins in any given algal genome have detectable sequence similarity to any defined domain in the Pfam database <sup>14</sup>.

Spatial organization of genes (i.e. their localization along the genome) provides an alternative and complementary source of information that is independent of primary sequence <sup>15</sup>. Genomic context-based methods, including gene neighborhoods, gene-order and gene-teams based methods, are a way to make use of this information <sup>12</sup>. These methods rely on orthology between genes and thus are subject to the above exposed limitations. Probably because these limitations, the few examples of genomic context-based AFP in eukaryotes are limited to a small proportion of the genes of the organism being considered <sup>29,41</sup>.

There is plenty of evidence pointing to the existence of distinctive patterns in the way in which functionally related genes distribute along eukaryotic genomes. If such patterns exist and are biologically relevant, it should be possible, at least in some cases, to predict the functions of a gene using as predictive feature its relative position with respect to other genes of known function in the same genome. As far as we know, here we have performed this task for the first time, using a new way to represent the information contained in these patterns: the Functional Landscape Arrays. This feature can be automatically extracted from any annotated genome and does not depend on orthology relations with other organisms.

Using FLAs as the only predictive feature we trained a set of hierarchical multi label classifiers that outperforms BLAST when predicting GO terms from the Biological Process Ontology in *H. sapiens*, *M. musculus* and *D. melanogaster* (see Figure 6). Our models also outperform BLAST when predicting GO terms from the Cellular Component Ontology in *H. sapiens* and *D. melanogaster*. With these trained classifiers we obtained thousands of new predicted associations between genes and GO terms in five eukaryotes, available at <http://gfpml.bnd.edu.uy>.

The relevance of our results stems from the fact that the performance of our models, assessed by standard metrics, shows that AFP exclusively based on

features derived from the relative location of genes can be successfully performed on eukaryotic genomes. Even though in AFP is usual to integrate multiple types of information, information derived from gene location is rarely taken into account. Furthermore, according to the CAFA organizers, new improvements in gene function prediction should be expected from the incorporation of new kinds of predictive features <sup>7</sup>. We believe that including FLAs as predictive feature could significantly improve the performance of AFP models.

Our results are interesting from another point of view. The existence in eukaryotes of distribution patterns of functionally related genes so well defined as to allow good AFP points to levels of organization thought to be exclusive of prokaryotic genomes and its characteristic operons <sup>42</sup>. Diament and Tuller performed a comparative study of the organization of several genomes, analyzing the location of functionally related genes. Their results revealed that the prokaryote *Escherichia coli* exhibits a higher level of genomic organization than the eukaryote *S. cerevisiae*, as one would expect given its operon-based genomic organization. But when considering a higher order of genomic organization, analyzing the co-localization of pairs of different functional gene groups, the authors found that the genome of *S. cerevisiae* is markedly more organized than that of *E. coli*. Our results are consistent with this trend. To estimate how far from randomness is the linear organization of different genomes we used the hF-max ratio, i.e. the ratio between the hF-max reached by the trained model and the hF-max reached by the random model. Table 2 and Figure 4 show that although the relationship between the complexity of the organism and its hF-max ratio is not linear, simpler organisms reach lower hF-max ratios than more complex organisms.

In sum, Functional Landscape Arrays have the potential to improve AFP, as they can be easily integrated into any model, can be automatically extracted from any annotated genome and are independent from sequence identity. To the best of our knowledge this is the first work in which only features derived from

the relative gene location of the genes within a genome are used to successfully predict gene function in eukaryotes.

**Competing interests:** The authors declare no competing interests.

**Acknowledgments:**

Funding: This work was supported by Agencia Nacional de Investigación e Innovación, Uruguay, [grant number FSDA\_1\_2017\_1\_14242]; Instituto de Investigaciones Biológicas “Clemente Estable”, MEC, Uruguay and Programa de Desarrollo de las Ciencias Básicas, Uruguay.

The experiments presented in this paper were carried out using ClusterUY (site: <https://cluster.uy>).

## References

1. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **46**, 2699 (2018).

2. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).
3. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res* **46**, D754-D761 (2018).
4. Rubin, A. F. & Green, P. Expression-based segmentation of the *Drosophila* genome. *BMC Genomics* **14**, 812 (2013).
5. Bernardes, J. S. & Pedreira, C. E. A review of protein function prediction under machine learning perspective. *Recent Pat Biotechnol* **7**, 122-141 (2013).
6. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nature Reviews Genetics* **16**, 321-332 (2015).
7. Zhou, N. *et al.* The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology* **20**, 244 (2019).
8. Zhao, Y. *et al.* A Literature Review of Gene Function Prediction by Modeling Gene Ontology. *Front Genet* **11**, (2020).
9. Bonetta, R. & Valentino, G. Machine learning techniques for protein function prediction. *Proteins* **88**, 397-413 (2020).
10. Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10**, 221-227 (2013).
11. Jiang, Y. *et al.* An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology* **17**, 184 (2016).
12. Shehu, A., Barbará, D. & Molloy, K. A Survey of Computational Methods for Protein Function Prediction. in *Big Data Analytics in Genomics* (ed. Wong, K.-C.) 225-298 (Springer International Publishing, 2016). doi:10.1007/978-3-319-41279-5\_7.
13. Duan, Z.-H., Hughes, B., Reichel, L., Perez, D. M. & Shi, T. The relationship between protein sequences and their gene ontology functions. *BMC Bioinformatics* **7**, S11 (2006).
14. Blaby-Haas, C. E. & Merchant, S. S. Comparative and Functional Algal Genomics. *Annu Rev Plant Biol* **70**, 605-638 (2019).

15. Ling, X., He, X. & Xin, D. Detecting gene clusters under evolutionary constraint in a large number of genomes. *Bioinformatics* **25**, 571-577 (2009).
16. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95**, 14863-14868 (1998).
17. Niehrs, C. & Pollet, N. Synexpression groups in eukaryotes. *Nature* **402**, 483-487 (1999).
18. Cohen, B. A., Mitra, R. D., Hughes, J. D. & Church, G. M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* **26**, 183-186 (2000).
19. Boutanaev, A. M., Kalmykova, A. I., Shevelyov, Y. Y. & Nurminsky, D. I. Large clusters of co-expressed genes in the Drosophila genome. *Nature* **420**, 666-669 (2002).
20. Hurst, L. D., Williams, E. J. B. & Pál, C. Natural selection promotes the conservation of linkage of co-expressed genes. *Trends Genet.* **18**, 604-606 (2002).
21. Lee, J. M. & Sonnhammer, E. L. L. Genomic Gene Clustering Analysis of Pathways in Eukaryotes. *Genome Res* **13**, 875-882 (2003).
22. Hurst, L. D., Pal, C. & Lercher, M. J. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet* **5**, 299-310 (2004).
23. Michalak, P. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics* **91**, 243-248 (2008).
24. Overbeek, R., Fonstein, M., D' Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96**, 2896-2901 (1999).
25. Huynen, M., Snel, B., Lathe, W. & Bork, P. Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences. *Genome Res* **10**, 1204-1210 (2000).

26. Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S. & Koonin, E. V. Genome Alignment, Evolution of Prokaryotic Genome Organization, and Prediction of Gene Function Using Genomic Context. *Genome Res.* **11**, 356-372 (2001).
27. Yanai, I., Mellor, J. C. & DeLisi, C. Identifying functional links between genes using conserved chromosomal proximity. *Trends Genet.* **18**, 176-179 (2002).
28. Zheng, Y., Roberts, R. J. & Kasif, S. Genomic functional annotation using co-evolution profiles of gene clusters. *Genome Biol.* **3**, RESEARCH0060 (2002).
29. Mihelčić, M., Šmuc, T. & Supek, F. Patterns of diverse gene functions in genomic neighborhoods predict gene function and phenotype. *Sci Rep* **9**, (2019).
30. Pazos Obregón, F. *et al.* Cluster Locator, online analysis and visualization of gene clustering. *Bioinformatics* **34**, 3377-3379 (2018).
31. Valentini, G. True path rule hierarchical ensembles for genome-wide gene function prediction. *IEEE/ACM Trans Comput Biol Bioinform* **8**, 832-847 (2011).
32. Boyle, E. I. *et al.* GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710-3715 (2004).
33. Tiirikka, T., Siermala, M. & Vihinen, M. Clustering of gene ontology terms in genomes. *Gene* **550**, 155-164 (2014).
34. Feng, S., Fu, P. & Zheng, W. A Hierarchical Multi-Label Classification Algorithm for Gene Function Prediction. *Algorithms* **10**, 138 (2017).
35. Feng, S., Fu, P. & Zheng, W. A hierarchical multi-label classification method based on neural networks for gene function prediction. *Biotechnology & Biotechnological Equipment* **32**, 1613-1621 (2018).
36. Breiman, L. Random Forests. *Machine Learning* **45**, 5-32 (2001).
37. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* **16**, 321-357 (2002).
38. Silla, C. N. & Freitas, A. A. A survey of hierarchical classification across different application domains. *Data Min Knowl Disc* **22**, 31-72 (2011).

39. Kiritchenko, S., Matwin, S., Nock, R. & Famili, A. F. Learning and Evaluation in the Presence of Class Hierarchies: Application to Text Categorization. in *Advances in Artificial Intelligence* (eds. Lamontagne, L. & Marchand, M.) 395-406 (Springer, 2006). doi:10.1007/11766247\_34.
40. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).
41. Foflonker, F. & Blaby-Haas, C. E. (ORCID:0000000215831291). Co-locality to co-functionality: Eukaryotic gene neighborhoods as a resource for function. *Molecular Biology and Evolution* (2020) doi:https://doi.org/10.1093/molbev/msaa221.
42. Diament, A. & Tuller, T. Three-dimensional Genomic Organization of Genes' Function in Eukaryotes. in *Evolutionary Biology: Convergent Evolution, Evolution of Complex Traits, Concepts and Methods* (ed. Pontarotti, P.) 233-252 (Springer International Publishing, 2016). doi:10.1007/978-3-319-41324-2\_14.

### **Author contributions**

FPO conceived and supervised the project, performed analyses and wrote the manuscript. DS performed and analyzed the experiments PS, GG, PY and RC discussed the results and corrected the manuscript. All authors approved the manuscript.