Computational Functional Annotation:
The Predictive Power of Different Data Sources

Itamar Borukhov* and Yossef Kliger
Computational R&D, Compugen Ltd.
26 Harokmim Street, Building D, Holon 5885849, Israel
*To whom correspondence should be addressed: itamarb@cgen.com

1. INTRODUCTION

In recent years, the CAFA project has established itself as an extensive effort to quantify and advance our ability to apply computational methods to predicting protein function annotations (1-2). In the two previous rounds and in the current ongoing one, teams of scientists have been developing computational methods and submitting predictions for sets of proteins that are missing some or all functional annotations. These predictions are tested later against new annotations that have accumulated since the submission deadline. The annotations primarily consist of the three GO annotation classes: Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). Most prediction methods rely heavily on sequence similarity and are complemented by other types of data and analyses.

In this study, we aim to understand better which data sources carry more predictive power for different classes of protein function annotations using datasets from the second CAFA round. Specifically, we quantify how the similarity in different properties, including gene expression profiles from GTEx, domain occurrence from Interpro, etc. compare with sequence similarity using Blast in predicting the different GO annotation classes (see, e.g., **Figures 1, 2**). In addition, we study which of these data sources adds more predictive power when combined with Blast. In our analysis, we consider two performance measures: (i) $F_{\max}$, the maximal value of $F = \frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$. The F-max measure is also implemented in the CAFA project and is a global measure dominated by the performance at intermediate precision and recall values. (ii) The second measure we consider is the area under the precision-recall curve up to a maximal recall value. The latter measure is focused on the top ranking (high precision) predictions. We discuss our results in light of the differences between the annotation classes: MF terms tend to be more specific, whereas BP terms tend to have a more abstract nature, and finally CC terms span a wide range of diverse proteins. For example, Interpro domains carry comparable predictive power to Blast for high scoring MF annotations but not for high scoring BP annotations (See **Figures 1, 2**).

2. FIGURES

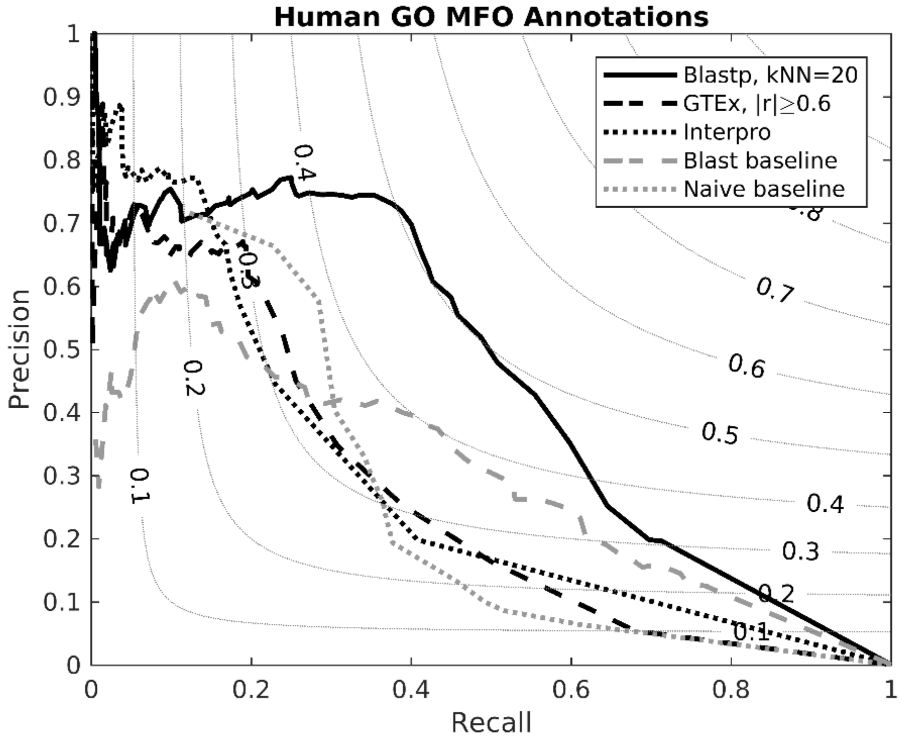

Figure 1: Recall-precision curves for predicting GO MFO annotations by different data sources and analyses: Naïve baseline (by term frequency), Blast baseline (by best hit), Blastp with kNN=20 nearest neighbors, GTEx gene expression with correlation coefficients of **|r|**≥0.6, and Interpro domain co-occurrence. The dotted contours represent equal values of $F = \frac{2\times\text{precision }\times\text{recall}}{\text{precision }+\text{recall}}$.

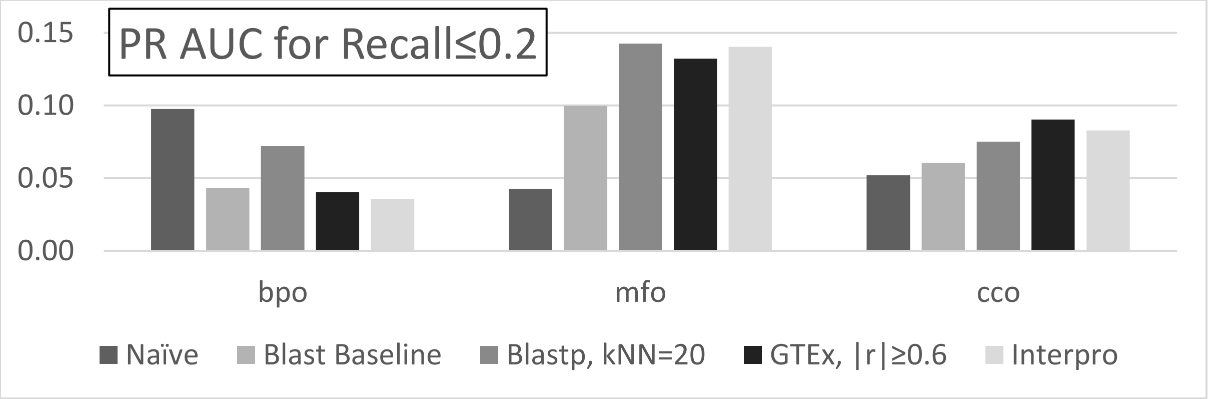

Figure 2: Performance by area under the precision-recall curve for Recall≤0.2 for the three GO annotation classes by different data sources and analysis: Naïve baseline (by term frequency), Blast baseline (by best hit), Blastp with kNN=20 nearest neighbors, GTEx gene expression with correlation coefficients of **|r|**≥0.6, and Interpro domain co-occurrence.

3. REFERENCES

1. Radivojac P, Clark WT, Oron, T. R., et al. 2013. A large-scale evaluation of computational protein function prediction, *Nature Methods* **10**:221–227

2. Jiang Y., Oron, T. R., Clark W. T. et al. 2016. An expanded evaluation of protein function prediction methods shows an improvement in accuracy, *Genome Biology* **17**: 184