Figuras de rendimiento

Gianlucca Colangelo

$$Score = \alpha \cdot Esp + \beta \cdot Cap + \gamma \cdot Sim \tag{1}$$

La especificidad y la similaridad están altamente correlacionadas (0.99 según el coeficiente de correlación de Pearson). De modo que nos podríamos deshacer de la primera y quedarnos solamente con la similaridad y la capitalidad para el modelo.

Par-métrica	**Coeficiente de correlación**
Similaridad / Especificidad	0.99
Similaridad / Capitalidad	0.22
Capitalidad / Especificidad	0.22

Cuadro 1: Se tomaron 500 HC al azar y se les calculó la tríada de métricas para el los 500 genes reales, y se evaluó cómo correlacionaban estas métricas a diferentes N. No se observaron cambios significativos a lo largo de N.

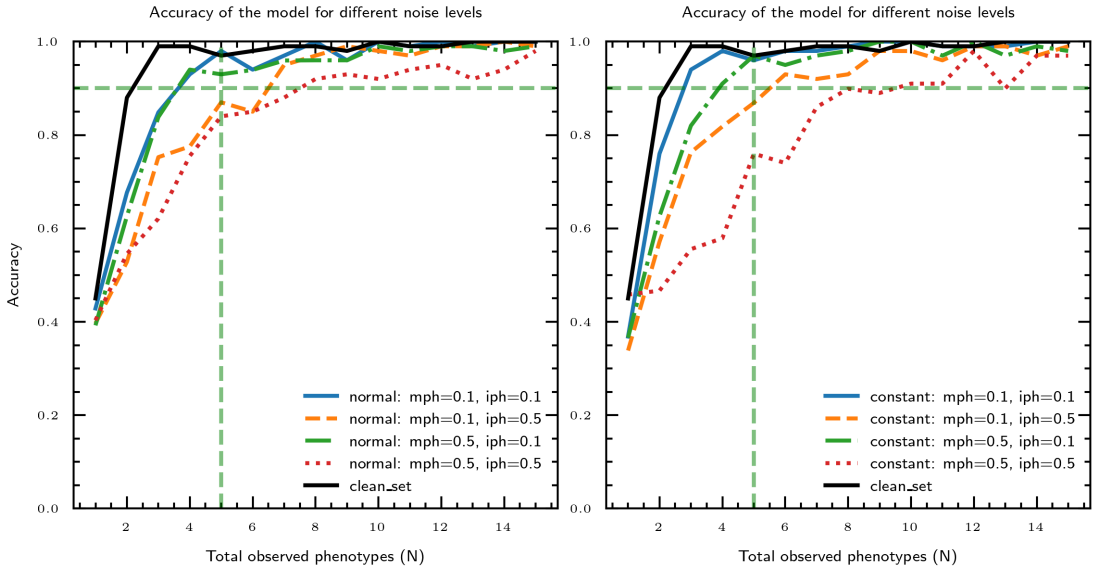
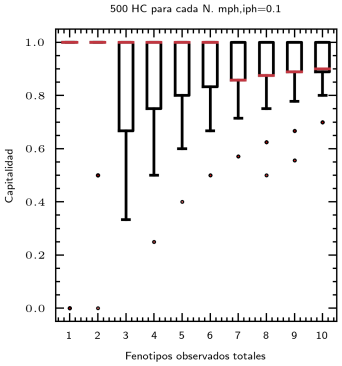
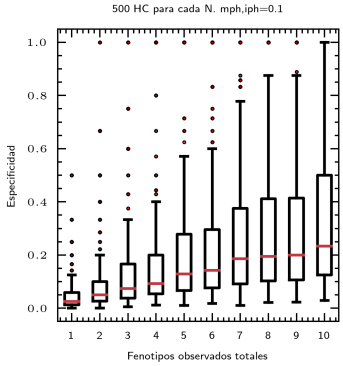


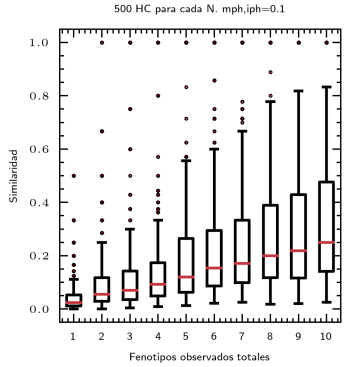
Figura 1: Rendimiento del modelo para diferentes tipos de ruidos con muestreo de 100 HC al azar para cada N. $Score = \alpha \cdot Esp + \beta \cdot Cap + \gamma \cdot Sim$



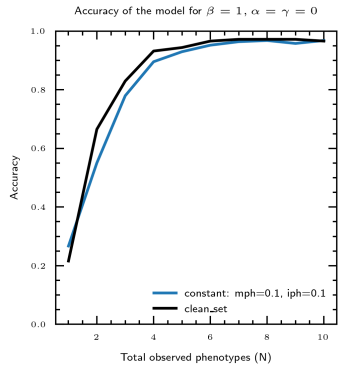
(a) Como se esperaba, se observa que la capitalidad tiende a 1 cuando el número de fenotipos observados totales es 1, ya que es más probable que los FO se encuentren en los fenotipos del gen.



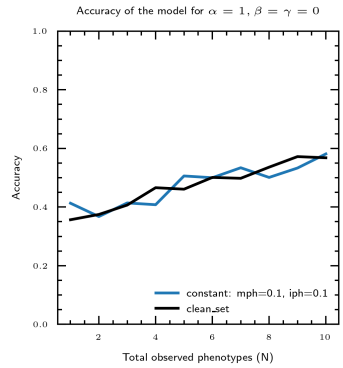
(b) Inversamente en cuanto a la especificidad. Es más probable que los fenotipos del gen tengan otros fenotipos además de los observados cuando estos son del ordend e 1 o 2.



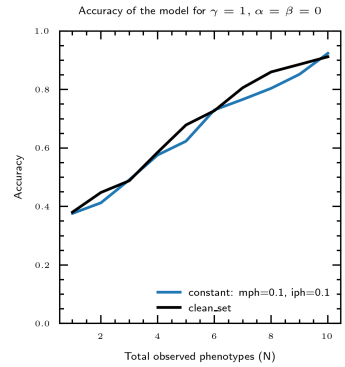
(c) Con la similaridad la tendencia aumenta linealmennte también. Esto se debe a que esta divide por el término de fenotipos únicos. Cuando aumenta el número de fenotipos observados totales, aumenta la intersección de los genes reales y disminuyen los fenotipos únicos.



(d) La capitalidad contribuye a un rendimiento mayor al 90 % cuando $N = 4$. Pero su rendimiento cae al 20 % cuando $N = 1$.



(e) La especificidad no parece aportar demasiado al modelo en comparación con la capitalidad. Salvo para $N = 1$ donde esta aporta un rendimiento del 40 %.



(f) La similaridad contribuye linealmente al rendimiento del modelo en función de N.

Figura 2: Para generar los gráficos de la primera fila se tomaron para cada N (fenotipos observados totales) 500 historias clínicas (HC) al azar (de aprox. 4900) y se calculó la métrica al gen real, para cada una de las HC. Como resultado tenemos un gráfico boxplot con la media en rojo y la varianza en la caja, para cada N. Luego los gráficos de la segunda fila fueron obtenidos definiendo un accuracy como los genes reales que quedaron en el top 10 con más score. Esto también se calculó para cada N, tomando 500 HC al azar.