# Rendimiento

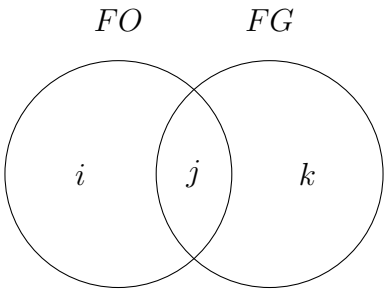



Figura 1: FO corresponde a Fenotipos Observados, con media en los sets de bitgenia y clinvar en una cantidad de 5. Mientras que Fenotipos del Gen, corresponde a los fenotipos registrados en OMIM para el gen candidato.

**Notar** que $FG$ es variable para cada gen candidato. Mientras que $FO$ es constante. Esto puede explicar luego los resultados con respecto a $k$.

Al principio habíamos definido:

$$C = \frac{j}{i+j} \qquad E = \frac{j}{k+j} \qquad S = \frac{j}{i+j+k} \tag{1}$$

Donde otras opciones eran:

$$\frac{i}{i+j}, \qquad \frac{k}{j+k} \tag{2}$$

Pero como notamos que cuando $j = 0$, nuestra métrica sería máxima, igual a 1, la modificamos del siguiente modo:

$$1 - \frac{i}{i+j}, \qquad 1 - \frac{k}{k+j} \tag{3}$$

Lo cual, si hacemos un poquito de algebra vemos que son equivalentes a $C$ y $E$ definidas en (1).

$$\frac{i+j}{i+j} - \frac{i}{i+j}, \qquad \frac{k+j}{k+j} - \frac{k}{k+j} \tag{4}$$

Sacamos factor común $\frac{1}{i+j}$, $\frac{1}{k+j}$ y tenemos que:

$$\frac{j}{i+j} = C, \qquad \frac{j}{k+j} = E \tag{5}$$

**Dicho eso:** Para empezar a definir métricas desde lo más simple, las preguntas que pensé que serían correctas hacerse son:

1. ¿Qué queremos **priorizar**, $i$,$j$ o $k$?
2. ¿Qué queremos **penalizar**, $i$,$j$ o $k$?

De lo cual concluí que por la naturaleza de nuestro problema, queremos **priorizar** $j$ y **penalizar** $i$ y $k$. Con lo cual ahora quedaría evaluar cómo debemos priorizar $j$ y cómo penalizar $i$ y $k$.

## Resultados

A continuación evaluo muchas combinaciones de métricas como habíamos hablado. También, para tener una manera cuantitativa de decidir la mejor, definí "**accumulated accuracy**", que es la suma acumulada del rendimiento para cada métrica a lo largo de N, normalizada. Accumulated accuracy igual a 1 significa que la métrica obtuvo un rendimiento del 100 % desde $N = 1$ hasta $N = 10$. Al final, elijo las 3 mayores.

### Lo más simple

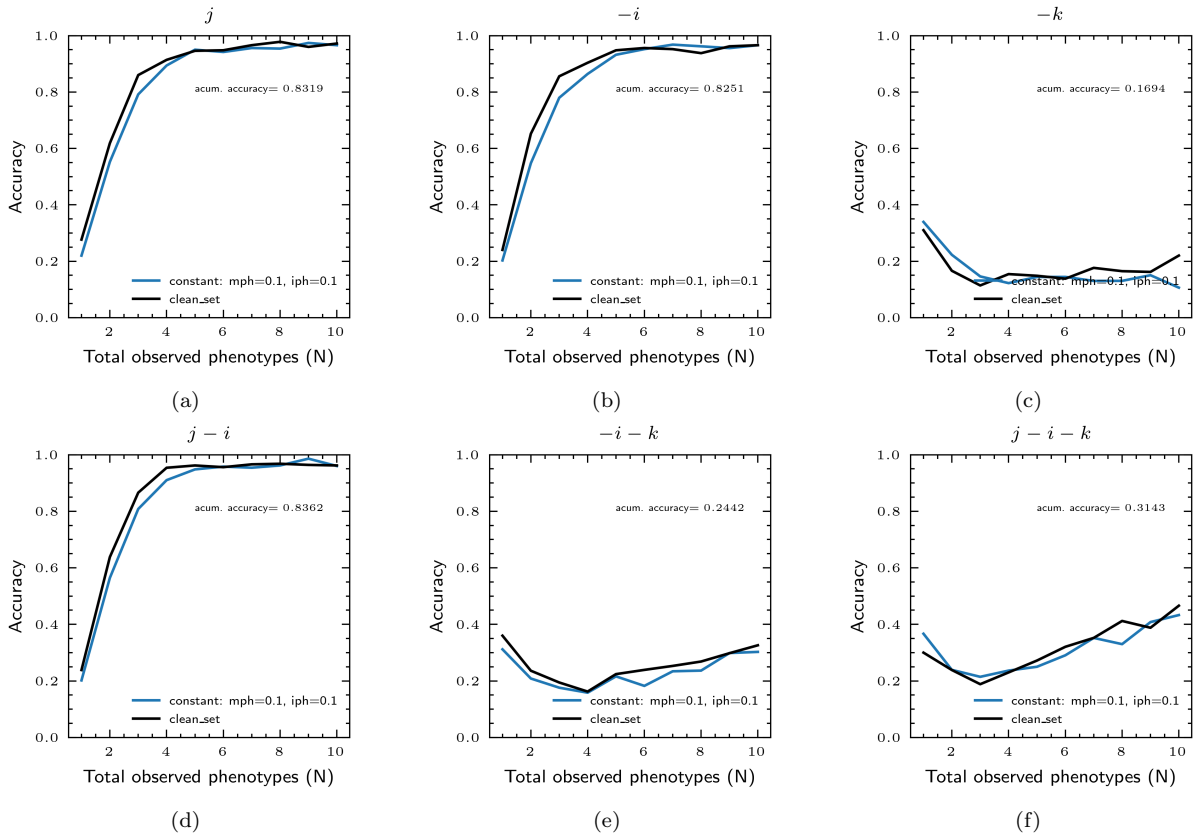



Figura 2: Rendimiento para cada métrica individual. Se puede observar que $-k$ no solo no contribuye al accuracy sino que lo perjudica. Por otro lado $-i$ y $j$ es esperable que den muy parecidas ya que son complementos en el conjunto $FO$.

### Ahora normalizando

La primera, segunda y tercera columna, corresponden a Capitalidad, Especificidad y Similaridad, respectivamente. Normalizadas primero linealmente, después exponencialmente y por último logarítmicamente, en cada fila. Hay casos que no están evaluados, varios los consideré innecesarios, si te parece que hay alguno que no debería obviarlo decime.

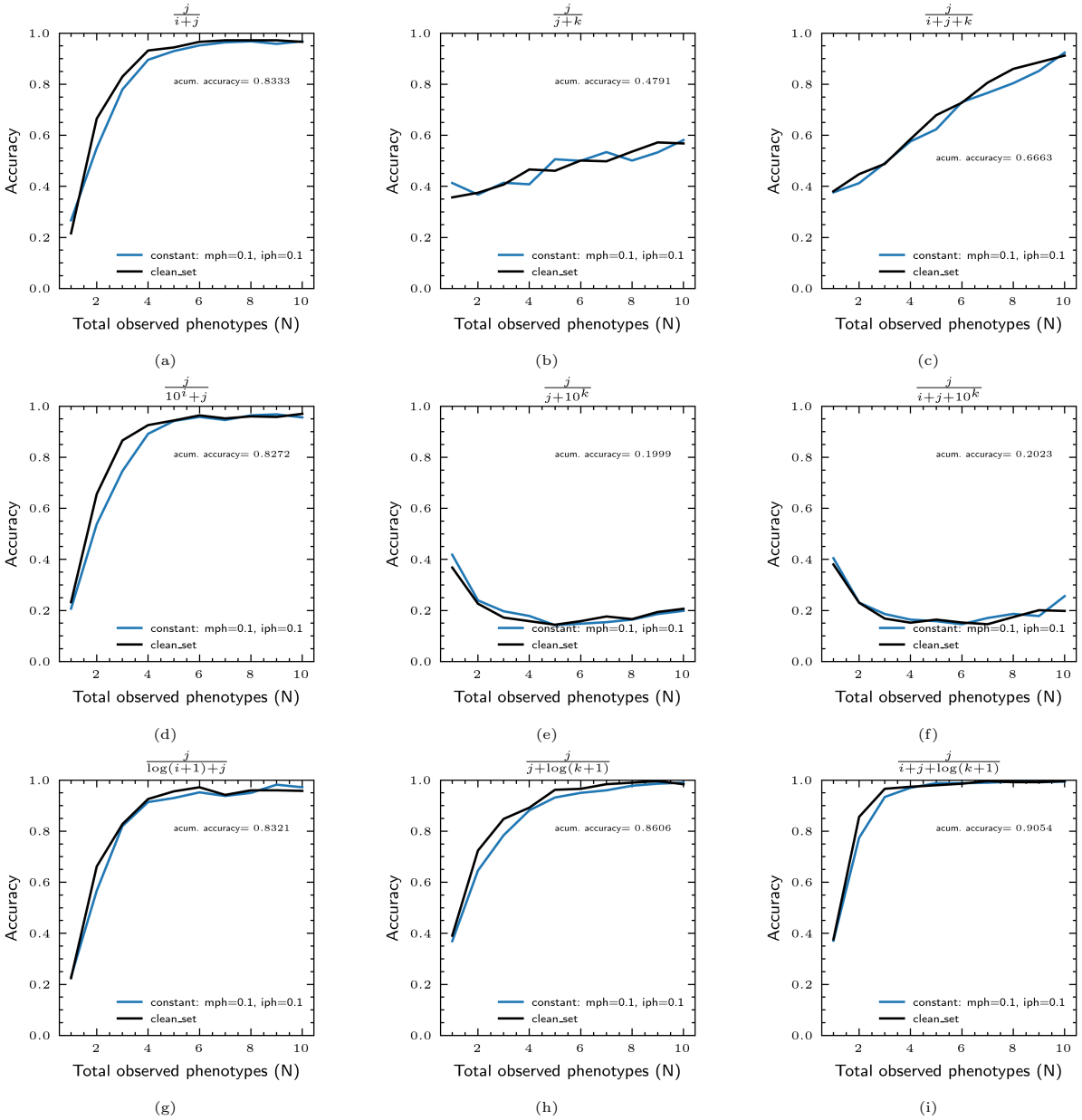
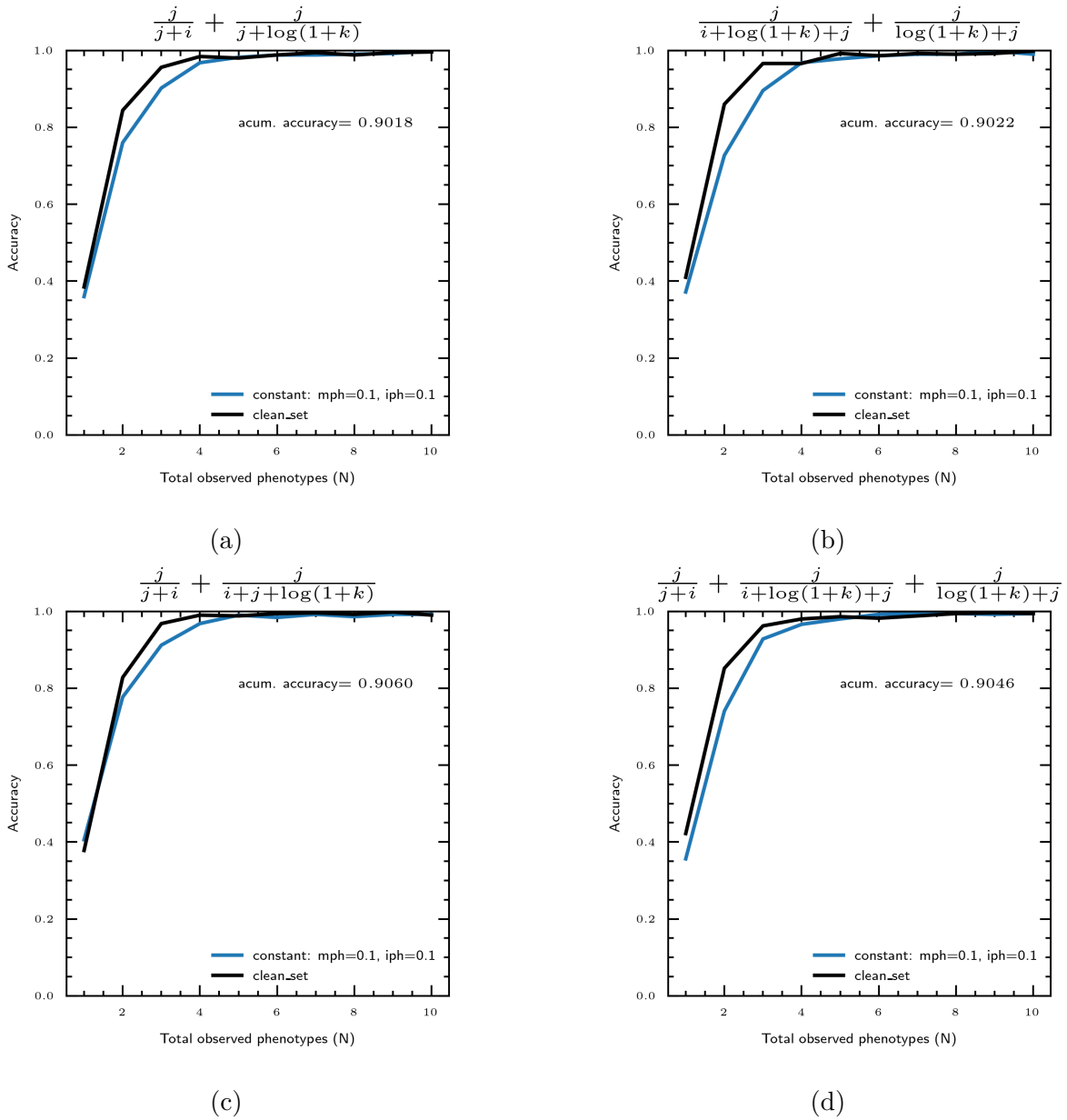



Figura 3: En la primera fila las métricas que habíamos definido inicialmente. La segunda fila con penalizaciones exponenciales en la normación y la tercera logarítmica.

Lo más interesante es que en la primera columna, tratar a $i$ lineal, exponencial o logarítmicamente no cambia en el rendimiento. Después, en el caso de la especificidad, tratar a $k$ logarítmicamente mejoró un montón. Y luego, en la similaridad también.

### Las tres más altas

Tomando las tres métricas que mayor acumulación de rendimiento tienen, procedí a probar combinaciones. No encontré ninguna que sea significativamente mejor a la de la Figura 3.i, con un accum. accuracy = 0,9054.




(a) (b)

(c) (d)

Figura 4

**Pensamientos** Alcanzado este punto, lo que quedaría probar, sería pesar las métricas con una restricción del tipo $\alpha + \beta + \gamma = 1$ y con algún algoritmo encontrar $\alpha, \beta, \gamma$ tal que maximice la función objetivo, que podría ser el rendimiento acumulado, actualmente en 0,905. Pero quizás en esta etapa eso sea hilar demasiado fino y sea mejor avanzar en otros aspectos. Otra cosa que pensé fue empezar a analizar cuáles son los casos para los que el modelo falla, ver qué propiedades tienen y ver si eso me pueda dar insights. Y por último, algo que hace bastante quiero probar, sería agregar una métrica más (que va por fuera del álgebra de conjuntos que estuvimos haciendo ahora), que consistiría básicamente

en convertir a $FO$ y $FG$ en vectores n-dimensionales, y (en lugar de intersección como venimos haciendo) medir la distancia coseno entre vectores, qué opinás?.

## Pesando fenotipos

Acá estuve haciendo algunos avances también, pero lo dejamos para la semana que viene mejor.