The paper deals with the longstanding and challenging problem of finding good and reliable negative examples specifically focusing on their selection for protein function prediction.

The paper is undoubtedly interesting and a step forward in this direction but it still needs some improvements and a revision.

Major points

1) The first thing is easy to address. Authors should convey better the message that selecting negative examples is not exactly the other side of the coin of the prediction of positive examples. In other words, it is not exactly providing "negative examples" by taking all of the GO terms minus positive annotations.
2) Selection of negative examples are given as a list of ranked proteins for a particular GO but I ask to:
    a. provide a score (raw score or better a p-value or something like that)
    b. provide "negative" GO for a protein and not only the GO with its list of "negative" proteins. This would be really helpful.
3) Authors show that the performance of SNOB (H. sapiens) is effective and performs better than other methods in predicting negative examples for more general GO categories i.e. GO terms with a low information content as they are close to the root of the GO graph and are more frequent. The performance decreases when GO terms are rare and subsequently more specific. What I suspect, on the contrary, is that methods have a general tendency to perform better, in any case, when GO terms are rare or have a low frequency if compared with those that are highly frequent. As an example one can check GO:0005515 "protein binding" and GO:0019901 "kinase binding" which are generic/highly frequent and more specific/less frequent respectively.
   Authors must explain what are the real effects they observe in the results when considering the strong biases in the frequencies of some GO terms which can vary a lot from very low to very high. It is not surprising that the number of erroneous negative examples increases proportionally with the higher frequency of the selected GO. Similarly, it is not surprising to perform well when the GO term is rare. This brings me back to what I said earlier, namely the need to have a score and not just an ordered list of proteins for a selected GO. This can help the user to choose on his own the desired threshold depending on the GO term (generic or specific) the user is interested in.
4) Finding a benchmark is extremely difficult and authors have had to deal with the same problem everyone has to face when assessing one's method. Nonetheless, I find a weak evidence reporting results for single GO terms and spend an entire paragraph. I suggest limiting this part and focus on general trends and what I ask in point 3)

Minor points

" ….. and (iii) using genes with annotations in sibling categories of the category …" (iii) what does this stand for ?

Some typos like "Rochhio" in First page of Introduction