Variations in the correlation of Gene Ontology annotations with sequence similarity over time.

Stuart Aitken stuart@inf.ed.ac.uk
Informatics, University of Edinburgh

The assignment of Gene Ontology (GO) terms to proteins has become an important method for characterising the function, process and cellular component of a protein such that comparisons can be made between proteins and their roles within and between species. Semantic similarity analysis is a means of combining the GO annotations of two gene products to derive a single similarity value by making use of the structure of the Gene Ontology and a database of annotations, from which probability values for shared GO terms can be calculated. Semantic similarity has previously been shown to correlate moderately with Blast score: The aim is to demonstrate a relationship between semantic similarity and the underlying biological reality. We examine the basis of semantic similarity from a probabilistic perspective. Two alternative views can be taken, i) that it is necessary to cover all annotations assigned to each gene (i.e. all genes have all assigned functions), or ii) that it is only necessary to cover some annotation made to each gene (i.e. alternative assignments can hold in different contexts). The similarity score can then be based on the minimum relative frequency of the terms that cover, or subsume, the annotations made to the gene products (the median or maximum are alternative choices), or the average can be calculated. However, we note that it is difficult to motivate the calculation of the average from a probabilistic perspective as it does not correspond to a joint or conditional probability. This yields four similarity measures, including one used in (Lord et al, 2003).

Following past investigations, we take annotations to human gene products from UniProt and analyse the correlation of semantic similarity scores with Blast score for proteins with a manually assigned GO terms. We perform this analysis at three different time points over the past two years. The four similarity metrics are tested for each of the three aspects of GO (molecular function, MF; biological process, BP; cellular component, CC).

In contrast with previous reports, we do not find that MF always has the best correlation with Blast score. Fig 1 shows a scatter plot of –ln(relative frequency) against ln(blastp bit score) for CC – the aspect of GO that showed the best correlation in the data from 2007. The two most consistent scores were those that covered some annotation made to each protein: min-of-min and ave-of-min. These measures give the same scores in the case where only one term is assigned to each protein, indicated by the black ∗. The remaining min-of-min scores (blue +) take the minimum of all minimum values for shared terms, while the remaining ave-of-min scores (red ×) average the minimum scores following the second assumption that a protein has all assigned terms. The figure shows that min-of-min points generally fall in the same area as those points where the measures agree, while the ave-of-min points have a low score even when Blast score (x axis) is high. R values for min-of-min range from 0.33-0.54, for ave-of-min R ranges from 0.11-0.43. Fig 2 shows a similar analysis for MF in 2005 but in this case similarity is plotted against ClustalW alignment. Previous results may have been influenced by the greater proportion of multiple term annotations made to BP and CC in earlier years, and the poor performance of ave-of-min on multiple terms.

The impact of this analysis is three-fold: i) that the standard method of averaging similarity scores is unsatisfactory from a theoretical perspective, and ii) less effective in practice than taking the minimum value, and finally iii) that the correlations exist for all aspects of GO, although they may be moderate and time-varying, however, the increasing numbers of annotations acquired over time do not appear to be detrimental to the similarity analysis.

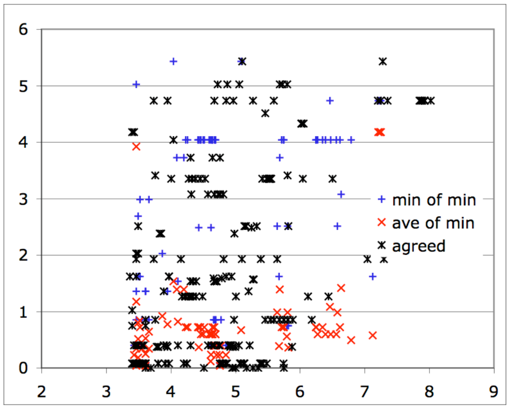

Fig 1. Similarity score vs ln(Blast) for CC 2007

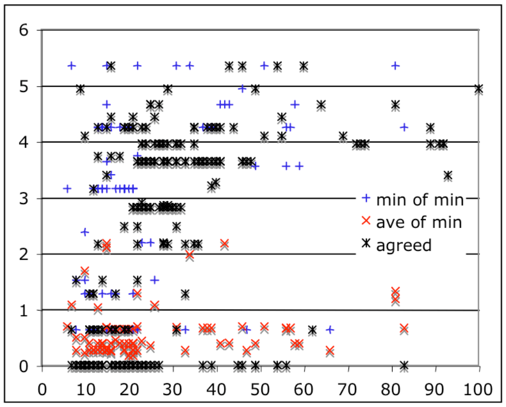

Fig 2. Similarity score vs ClustalW for MF 2005

Lord, P.W., Stevens, R.D., Brass, A. and Goble, C. (2003) *Proc. PSB* 8:601-612