Genome analysis

GOSLING: a rule-based protein annotator using BLAST and GO
Craig E. Jones1,2,*, Julian Schwerdt1,3, Tessa Arwen Bretag2, Ute Baumann1 and Alfred L. Brown2
1Australian Centre for Plant Functional Genomics, Waite Campus, University of Adelaide, South Australia 5064, 2School of Computer Science, University of Adelaide and 3South Australian Partnership for Advanced Computing, University of Adelaide, South Australia 5001

Received on May 23, 2008; revised on September 1, 2008; accepted on September 10, 2008
Advance Access publication September 16, 2008
Associate Editor: Dmitrij Frishman

ABSTRACT
Summary: GOSLING is a web-based protein function annotator that uses a decision tree-derived rule set to quickly predict Gene Ontology terms for a protein. A score is assigned to each term prediction that is indicative of the accuracy of the prediction. Due to its speed and accuracy GOSLING is ideally suited for high-throughput annotation tasks.
Availability: https://www.sapac.edu.au/gosling
Contact: craig@cs.adelaide.edu.au

1 INTRODUCTION
The exponential increase in sequence data worldwide has made it impossible to assign functional information to every sequence manually. To address this, a significant research effort is being made in the development of new automated approaches to annotating genes with functional information. Accordingly, because of the widespread use of the Gene Ontology, or GO (Ashburner et al., 2000), to describe gene function, there has been a growing interest in the development of electronic annotators that automatically predict GO terms. Perhaps the most commonly utilized method for predicting GO terms for a sequence is based on sequence similarity to previously annotated sequences. This method (referred to as ISS, Inferred by Sequence Similarity) is the most popular annotation approach accounting for 42% of all curated GO sequence database annotations (http://archive.geneontology.org/full/2006-03-01/).

Here we describe GOSLING (GO similarity listing using information graphs), a web-based tool for the annotation of protein sequences that predicts the accuracy of potential term annotations associated with BLAST-matched sequences present in the GO sequence database (http://archive.geneontology.org/full/). A number of other annotators have already been developed including GOplet (Hennig et al., 2003), GOFigure (Khan et al., 2003), GOfetch (Martin et al., 2004) and GOpet (Vinayagam et al., 2006). GOSLING makes predictions rapidly, requiring a few seconds for several hundred sequences, making it ideal for high-throughput sequencing projects. GOSLING enhances the utility of sequence information by extracting functional knowledge with a lower annotation error rate than occurs during standard annotation methods. For this reason, GOSLING can be used to verify hypothetical or predicted functional annotations of protein sequences.

2 ALGORITHM DEVELOPMENT
The March 1, 2006 GOSeqLite database release was downloaded and installed (http://archive.geneontology.org/full/2006-03-01/). This database contains only curated GO term annotations for protein sequences. Predicting sequence annotations using rules derived from cases that were annotated using only sequence similarity, results in poorer prediction precision (Jones et al., 2007). For this reason, sequences and their annotations were included in the test sample where sequence annotations were assigned with evidence that was not inferred from sequence similarity (non-ISS). Using this inclusion criterion, a total of 59 251 sequences and 182 828 annotations were selected for prediction accuracy benchmarking. A 10-fold cross-validation and pruning were used to avoid overfitting (Russel and Norvig, 1995). Matching sequences were identified by the NCBI blastp application (Altschul et al., 1997) at an expect value cutoff of $1 \times 10^{-10}$. Previous work (data not shown) had demonstrated that including more than five best matching sequences did not improve prediction precision. As such, the five best matching sequences for each query, and their associated non-ISS GO term annotations ($N = 417 834$), were selected. Data were examined for attributes and the following were identified: term depth, term usage count, term found before, term ontology, evidence code, BLAST expect value, BLAST score bits, result rank exclusive and result rank inclusive. GO term duplication among matching-sequence annotations were processed such that each unique GO term only had a single instance, and duplicates were reflected in the ‘term found before’ value. As the correct set of annotations was known for each query sequence, each potential term annotation was also flagged as correct or incorrect. Statistical models were constructed relating term annotation attributes to the probability that the annotation was correct using decision tree analysis. SPSS (SPSS Inc, 2004) was employed to create Classification and Regression Trees on a training set of 417 834 term annotations. A C-score (an indicator of the likelihood that an annotation is correct) was calculated for all merged terminal nodes, where the C-score was defined as the proportion of correct cases for that node. This model was used by GOSLING for its classification algorithm logic.

3 EVALUATION
Evaluating the output of functional annotation predictors is difficult due to the scarcity of unbiased test data (Martin et al., 2004). Our approach is to use a benchmark method that assigns GO terms to a
protein that are associated with the highest scoring BLAST-matching sequence under several conditions. We refer to the benchmark method as Best-BLAST (Jones et al., 2005). Best-BLAST was chosen because it closely resembles how biologists might determine the function of novel lab-derived proteins by simply using BLAST to find the most similar GO annotated protein. The precision of annotations was defined as being the proportion of correct predicted terms for all predicted terms.

3.1 Training set

Training set cases were used as input initially to verify the C-score metric and provide an optimistic benchmark against Best-BLAST. Linear regression showed that an association between C-score and the proportion of correct term annotations was significant ($P < 0.001$, $r^2 = 0.98$). Best-BLAST resulted in a precision of 0.51 ($N = 48,040$ annotations). Comparatively, potential term annotations with a C-score $\geq 0.5$ had a precision of 0.69. As such, GOSLING was 35% more precise than Best-BLAST at predicting the 182,882 annotations associated with the March 2006 training corpus.

3.2 Non-ISS test set

A more conservative evaluation of the performance of GOSLING compared with Best-BLAST was developed by using only high-quality curated annotations. UniProt annotations are considered high quality due to the exhaustive curation methodology undertaken by human experts during their assignment (Apweiler et al., 2004). The June 27, 2006 UniProt release (13,296 manually annotated protein sequences with non-ISS GO term annotations) was downloaded as test data. A total of 12,913 sequences had significant BLAST matches. GOSLING was executed to assign a C-score to potential term annotations for each query protein sequence. A total of 134,609 annotations were scored. To prevent bias introduced by sequences being present in both the UniProt and GOSeqLite databases matches with an expect value of 0 were excluded. This artificially decreased the absolute precision of these methods making them suitable for relative comparisons only. Potential term annotations with a minimum C-score value of 0.5 were selected as putative annotations.

The resulting precision of these annotations was 0.55. This was compared with the precision generated by using a Best-BLAST annotation method, which was found to have a precision of 0.31. As such, GOSLING was 15% more precise than Best-BLAST.

In summary, we have shown that GOSLING has a 15–35% greater precision than Best-BLAST. GOSLING also provides a ranking of the expected relevance of terms that fall outside of the C-score cutoff, so that human curators may include additional GO terms that seem relevant. In this way it aims to provide a tool for curators, as well as providing a completely automated method. However, it is worth noting that predicting protein function based on a database of examples will be biased by the species composition and research focus of the cases used.

4 GOSLING WEB APPLICATION

The February 10, 2008 GoSeqLite database was downloaded, and a new set of rules were generated using the process described above. To ensure the utility of the new rule set for human curation tasks the outputted predictions for 100 protein sequences of known function were manually examined. The updated model was then adopted and incorporated into the GOSLING engine. The GOSLING application and source code are available at https://www.sapac.edu.au/gosling. GOSLING is used by entering FASTA formatted protein sequence data manually or by specifying a file for upload. Sequences are then submitted to a high-performance cluster for BLAST search against a custom database of non-sequence similarity-based (non-ISS) annotated sequences from the GoSeqLite database. Non-ISS derived GO term annotations associated with matching sequences are selected as term predictions. Predictions are assigned a C-score based on GO term and BLAST match attributes. The C-score is an estimate of the probability that the term annotation is correct. When complete, all GO terms associated with similar non-ISS annotated sequences are displayed in descending order of C-score. C-scores range between 0 and 1, with terms assigned higher C-scores considered to be more reliable functional predictions.

5 CONCLUSION

GOSLING predicts the function of protein sequence data with use of a decision tree-derived rule set. GO terms are predicted for novel protein sequences, and assigned a corresponding C-score which indicates the likelihood that the prediction is correct. We compared the accuracy of GOSLING annotation predictions against those produced by the commonly used Best-BLAST annotation method. For non-ISS annotated test sequences, potential term annotations receiving a C-score of 0.5 or greater were more precise than term annotations assigned by a Best-BLAST approach. GOSLING bases predictions on curated sequence annotations not inferred from sequence similarity (i.e. curated non-ISS annotations) as these are likely to be the least error prone. Due to the fact that GOSLING uses a relatively small number of rules, GOSLING is comparatively fast, enabling it to make predictions in a matter of seconds. GOSLING is available online at https://www.sapac.edu.au/gosling as a web-based or downloadable standalone application.

ACKNOWLEDGEMENTS

The authors wish to thank the Australian Centre for Plant Functional Genomics and eResearch SA for support during this research project.

Conflict of Interest: none declared.

REFERENCES


