GORetriever: a novel Gene Ontology annotation tool based on semantic similaritiy for knowledge discovery in database

Fontana P (1), De Mattè L (1), Cestaro A (1), Segala C (1), Velasco R (1), Toppo S (2)

(1) IASMA Istituto Agrario di S. Michele all'Adidge (trento)
(2) Dipartimento di Chimica Biologica Università degli Studi di Padova

Motivation

Over the years, biological databases have grown at a spasmodic rate and we have assisted at an exponential increase of the available amount of data. The biological knowledge, associated to a particular sequence, is usually expressed in natural language and stored as free text. Most of the information is unstructured and does not follow strict semantic rules. End users can easily understand this human readable format but the same knowledge cannot be managed and caught by a computer program. The Gene Ontology (GO) (1) effort goes in this direction providing a structured vocabulary where each term is described as a father-child relationship and multiple inheritances are allowed. In this framework protein functions are represented by a DAG (Directed Acyclic Graph) starting from the root, consisting of general terms, to the leafs containing different levels of detailed descriptions. Such an ordered infrastructure makes feasible to infer and measure semantic similarities of distant or different concepts simply looking at the information content they share.

Methods

Our method is an approach to automatically annotate sequences based on retrieved GO terms. The starting list of GO terms to evaluate may be obtained, for instance, by a simple similarity search of the query sequence against a database of GO annotated proteins. The GO hits are processed in order to reconstruct all of the possible paths that lead to the root node. During the recursive process each node is scored adding the weights of the nodes encountered during the path reconstruction. As a result we obtain a trimmed GO graph consisting only of the terms found in the database search: for each term we keep track of its occurrence and of its cumulative score. The algorithm calculates the Z score of the cumulative score obtained for each node. The path, including the nodes with the highest weights, is extracted. Due to the additive property used to weight the nodes, only generic annotation terms are discriminated efficiently. These nodes are near the root node and therefore they are highly frequent. To solve this problem a different measure has been used to get a good tradeoff between detail information and statistical significance. The nodes belonging to the most probable selected starting path, may contain too many GO terms. They are, then, grouped on the basis of their Information Content (IC, based on the frequency of each term) and their semantic distance calculated applying the Lin (2) formula that quantify the amount of information shared. This clustering criterion of similar terms allows to restrict the searching space of correct annotations. The remaining term hits are then ranked efficiently using two statistical scores and an entropy based measure: "Internal Confidence" (InC), "Absolute Confidence" (AC) and Theil Index (TI). The InC and AC scoring methods have been specifically developed to assess the statistical significance of the retrieved hits and are both based on non-cumulative node weights divided by either cumulative root node weight (InC) or by the maximal theoretical weight (AC). Theil index (TI) (3) is derived from Shannon's measure of information entropy and it is applied to measure the inequality of score distribution over the trimmed GO graph. The final retrieval step is based on ranking GO terms with the highest score and information content (IC).

Results

GORetriever has been benchmarked using SwissProt Release 48.7. We randomly extracted from the database four sets of 5000 sequences sharing less than 10% of sequences one another. We tried to recover the original GO annotation using a simple BLAST search against SwissProt Release 48.7 from which test set sequences have been removed. Each GO term has been ranked according to its AC and InC score while TI index has been considered as an entropy measure of the GO distribution in the graph. Three different categories have been taken into account to evaluate GORetriever performance: "exact" matches are GO terms identical to the original ones, "similar" matches are GO

terms that are very close to the original ones (i.e father-child relationship), "mismatches" hits are the remaining cases. Almost the totality of the sequences in each test set have been annotated and the amount of correct hits that GORetriever has assigned is similar in each test set showing both high sensitivity and selectivity. On average 91% of the whole hits are "exact" matches, 3% are "similar" matches and 6% are "mismatches".

Contact email: paolo.fontana@iasma.it

References

1. Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. Nucleic Acids Res. 1;34(Database issue):D322-6, 2006.
2. Dekang Lin. An Information-Theoretic definition of similarity. Proceedings of the Fifteenth International Conference on Machine Learning, pages 296-304, 1998.
3. Theil Henri. The measurement of inequality by component of income. Economics Letter, 2, 1979