MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Protein Annotation with GO Codes

Elena Catona[b], Padmini Srinivasan[ab], W. Nick Street[b]

[a]*School of Library and Information Science, The University of Iowa, Iowa City, USA*
[b]*Department of Management Sciences, The University of Iowa, Iowa City, USA*

Abstract

In this paper our goal is to present results from experiments with assigning Gene Ontology (GO) codes to a subset of Swiss-Prot database pertaining to human proteins using a supervised classification method. Our approach is to first classify documents referenced in the Swiss-Prot subset as relevant to proteins with codes, then annotate each protein with a subset of codes assigned to its relevant document. We classified a subset of 6,295 proteins with all GO codes that are in the subset (2,215 codes) and obtained F-measures of 82% for proteins with cellular component codes, 46% for molecular function codes, and 62% for biological process codes.

Introduction

Our research objective is to explore a supervised classification approach for automatically annotating proteins with codes from the Gene Ontology (GO) nomenclature. Our system takes as input a protein and a set of relevant documents and annotates proteins with one or multiple GO codes. Documents relevant to proteins could be obtained by key phrase search for proteins or co-occurrence with proteins.

GO is a systematic and standardized nomenclature for the description of genes and gene products in organisms. GO consists of three separate ontologies describing molecular functions (F), biological processes (P), and cellular components (C). We classify a subset of SWISS-PROT pertaining to human proteins with all GO codes in the subset. SWISS-PROT is a protein knowledgebase offering high quality annotation and direct links to specialized databases with minimum redundancy. The particular file that is the source of our dataset contains the GO Annotations (Human) produced by EBI (European Bioinformatics Institute).

Methodology

Our approach for automatically annotating proteins is two-staged. First we classify documents with codes with binary Naïve Bayes classifiers and then we annotate proteins with codes via their set of relevant documents. We perform separate experiments on the three ontologies of GO. Our methodology within each experiment is that of a three times tenfold-cross-validation design for each hierarchy, where the folding is done on the sets of proteins with function, component and process codes, respectively. When building a single binary document classifier the positive examples are the training set documents that are assigned that code. The negative set is the union of positive documents for all other codes in the hierarchy except for the documents that overlap with the positive set of documents.

Proteins are assigned codes using 'majority votes', where a 'vote' is given by the number of documents in the proteins's set that was assigned the code. A protein is assigned all codes that the classifier assign to the protein's document set that have a number of 'votes' at least equal to the margin. The margin is the difference between maximum and minimum number of votes.

Results

The best Precision and Recall values were obtained for proteins with cellular codes, .85 and .75 respectively. For biological process codes Precision is .76 and Recall .52, and for mollecular function codes .57 and .43.