A gene annotation tool based on sequence similarity and biomedical text.

Student ID : 47-56912
Name : Daisuke Yamaguchi
Thesis Supervisor : Prof. Toshihisa Takagi

1. Introduction

In recent years, a substantial number of prokaryotic genomes were completely sequenced and now additional prokaryotic genome sequencing projects are ongoing. On the other hand, with the development of new sequencer, direct cloning approach without cultivation was made possible and studies of metagenome such as the intestinal bacteria flora have been spotlighted. Although many genes have been found in these studies, most of them have not been functionally annotated yet. To address this problem, many computational methods have been proposed such as homology-based annotation transfer, structure prediction, integrative functional genomics, and so on. The most major method for assigning an annotation to a newly sequenced gene is to utilize existing information of its similar sequences. It does not, however, work well since its similar sequences often have a variety of annotations or lack of reliable annotations. Some of similar sequence functions are varied in biomedical literature and are not provided for sequence annotation. The exploitation of biomedical literature for sequence annotation is a crucial subject.

In this study, we propose a method and implemented it as an annotation system that can be applied to these sequences by combining a homology-based technique and biomedical text.

2 Method

Our system is composed of three parts, the words used for annotation are selected in 2.1 and the gene name recognition and selection of feature words are done in 2.2 in advance, the annotation for query sequence is interactively calculated in 2.3

2.1 Selection of words for annotation

To consider various functional words without including meaningless words, likely insignificant words were predicted using mutual information between each word and each MeSH term based on their co-occurred frequencies in MEDLINE abstracts. When the highest MI among all MIs with each MeSH term is lower than a threshold, the word is regarded as a meaningless word. The threshold is determined by a preliminary study. Words predicted as insignificant are removed from all words with high tf*idf.

2.2 Recognition of gene names and selection of feature words

First, we recognize prokaryotic gene names in MEDLINE abstracts with MeSH terms related to prokaryotes automatically. Second, we extract feature words for

each gene from the abstract set containing the gene name. Third, we store these feature words in a database with the gene names associated with their sequences.

2.3 An automated annotation procedure

First, using a sequence that a user wants to annotate as the query, he/she does a BLAST search against the database. Our system combines the query sequence with the feature words of the retrieved sequences, which are candidate words expressing functions of the query. Then, appropriate words are selected. The meaningless words calculated in 2.1 are not used in this annotation.

To identify appropriate feature words, which we assume are frequently used in the descriptions of retrieved sequences or specifically used in the descriptions of highly ranked sequences,

In our method, appearance frequencies and ranks in a BLAST hit-list are used for scoring feature words. The words with scores higher than a threshold would be taken as appropriate.

3 Results and Discussions

To investigate the performance of our gene name recognition step, we extracted 32,732 of gene-PubMed ID relations as a gold standard from Entrez Gene (gene2pubmed) and used for the calculation of recall. For the evaluation of precision, 100 abstracts are randomly selected from all abstracts where prokaryotic gene names are recognized our method. As a result, the recall was 0.54 and precision was 0.91.

In addition, we surveyed how many candidate feature words were covered by existing ontology/controlled vocabularies (i.e., GO, COGs, UMLS and TIGR roles). Since the valuable words for annotations were distributed within wide tf*idf range, we took 600 representative words and evaluated them by hand. As a result 202 of valuable words were contained in them, and their 67.8% of words were not registered in those vocabularies. Therefore free words additions to the controlled vocabularies are utilized in our annotation system.

To assess the result of our annotation, we made a gold standard from data of TIGR CMR, and evaluated the performance. When we evaluated the system using 20 genes of gold standard, most of resulting annotations (words/terms) were occupied with meaningless words or terms without word filtering in 2.1. Its precision was 0.110-0.189. Finally, by word filtering in 2.1 are used, the precision of 0.217-0.446 was achieved. The precision of words for annotation was improved about two times better by the filtering.

This system is expected to be useful for annotations of metagenome sequences or fragments of genes.