

# Specificity-based Ranking Algorithm for Finding Associated Terms from MEDLINE Abstracts

Yeondae Kwon<sup>1</sup>      Hideaki Sugawara<sup>1</sup>  
 yekwon@lab.nig.ac.jp      hsugawara@genes.nig.ac.jp

<sup>1</sup> Laboratory for Research and Development of Biological Databases, CIB-DDBJ Center, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

## Abstract

In this paper, we propose a ranking algorithm that focuses on specificity to a particular disease. In a specificity-based ranking method, a gene that causes a given disease but does not cause other diseases is ranked at the top. This is helpful in avoiding side effects in drug developments.

**Keywords:** ranking algorithm, specificity, co-occurrence, gene prioritization

## 1 Introduction

There are a lot of interests on extracting associations between diseases and genes from literatures such as MEDLINE abstracts. For a given disease, a search engine returns a ranked list of candidate genes according to some criterion. In this paper, we propose a ranking algorithm that focuses on specificity to a particular disease. A specificity-based ranking method should rank a gene that causes a given disease but does not cause other diseases at the top. This is important for drug developments because users can find relevant genes that do not have side effects quickly. For example, if some gene is specific to Alzheimer's disease, a chemical compound relevant to the gene may be useful for drug developments for Alzheimer's disease. In this paper, we describe a specificity-based baseline algorithm using term dictionaries and co-occurrence data of terms in MEDLINE abstracts and discuss future directions.

## 2 Specificity-based Ranking Algorithm

Strength of relevance between biological terms can be measured using various vocabularies such as eVOC, MESH, and Gene Ontology [1], but first we begin by using co-occurrence data of terms in MEDLINE abstracts. Incorporation of other vocabularies is the future work.

### 2.1 Measuring Specificity

One of criteria indicating strong association between disease and gene is the number of MEDLINE abstracts in which both of the two terms appear. However, if the gene is related to another disease, in other words, co-occurrence of the gene and another disease is also frequent, drug development based on the gene may lead to overlook its side effect to the other disease.

So, we introduce the notion of specificity with respect to a disease, which is measured by whether there is a document in which a gene and the disease co-occur and how less the gene co-occurs with other diseases. Specificity-based ranking algorithm is formally defined as follows. Let  $G(d)$  be the set of genes that co-occurs with disease  $d$  in some document and  $D(g)$  be the set of diseases that co-occurs with gene  $g$  in some document. When prioritize genes for a given disease  $d$ , sort  $g \in G(d)$  according to the cardinality of  $D(g)$  in an ascending order.

The definitions of co-occurrence frequency and specificity correspond to the notions of term frequency and inverted document frequency in document retrieval [2], respectively. These two criteria can be combined as one score like tf-idf method.

## 2.2 Synonyms

Each gene/disease has its synonyms and they appear with various names in MEDLINE abstracts. A primary gene and its synonyms should be identified as the same gene/disease. Thus, we incorporate the following synonym processing. Let  $S(d), S(g)$  be the sets of synonyms of disease  $d$  and gene  $g$ , respectively. For a given disease  $d$ , first take the union of all  $G(\hat{d})$  for each  $\hat{d} \in S(d)$ , and then obtain  $D(\hat{g})$  for each  $\hat{g} \in S(g)$  for each  $g \in G(\hat{d})$ . Then count distinct diseases in  $D(\hat{g})$  for each gene group.

## 2.3 System Implementation

We use two dictionaries, NCBI Gene and OMIM databases. A group of genes that have a same Entrez GeneID are synonyms. Using terms in these dictionaries, construct terms $\times$ documents tables for gene and disease which represent term occurrence in MEDLINE abstracts. Then count the number of documents for each combination of gene and disease by joining the terms $\times$ documents tables by PMID. Specificity can be computed by the above tables.

## 3 Discussions

There are other factors which have the possibility to contribute to specificity. Among them, we plan to incorporate gene families into a score. For example, if genes  $g_1$  and  $g_2$  belong to a same family and  $g_1$  has an association with disease  $d$ , then  $g_2$  is assumed to have a weak association with  $d$  even if  $g_2$  and  $d$  do not co-occur in any documents. We will decide whether given two genes belong to a same family by finding common prefixes after normalization of genenames.

Several factors should be combined as one specificity score. If a benchmark dataset is available, machine learning techniques such as neural network can be applied for determining weights of factors instead of pre-defining them. Refinement of co-occurrence data using some natural language processing technique is also a future task.

The precision of the specificity-based ranking algorithm will be evaluated by comparing other prioritization schemes (e.g., co-occurrence, a combination of co-occurrence and specificity, and expectation value [3]) using some dataset such as the Endeavour system [4], which prioritizes over 600 genes in disease dataset.

Furthermore, we plan to add chemical compounds as biological terms and provide a workflow system which helps comprehensive analysis of genes, diseases, and chemical compounds.

## References

- [1] Yu, S., Vooren, S., Tranchevent, L., Moor, B., and Moreau, Y., Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining, *J. Bioinformatics*, 24(16):i119–i125, 2008.
- [2] Baeza-Yates, R. and Ribeiro-Neto, B., *Modern Information Retrieval*, Addison-Wesley, 1999.
- [3] Tsuruoka, Y., Tsujii, J., and Ananiadou, S., FACTA: a text search engine for finding associated biomedical concepts, *J. Bioinformatics*, doi:10.1093, 2008 (to appear).
- [4] Aerts, S., et al., Gene prioritization through genomic data fusion, *Nat. Biotechnol.*, 24, 537–544, 2006.