PUBMED ABSTRACT PROCESSING FOR PROTEIN FUNCTION PREDICTION

A thesis submitted by
LI ZHIHUI (U042268W)
in partial fulfilment for the
Degree of Bachelor of Science
with Honours in
Computational Biology

Supervisor: Professor: Wong Lim Soon
Co-supervisor: Associate Professor Choi Kwok Pui

NATIONAL UNIVERSITY OF SINGAPORE
2007/2008

ACKNOWLEDGEMENTS

My deepest gratitude goes to my supervisors, Professor Wong Lim Soon and Associate Professor Choi Kwok Pui, for the advice and guidance provided during the course of the project. I would also like to extend my gratitude towards Chua Hon Nian, who has provided immense help and advice on issues pertaining to the program in use and allow the proficient use of the program. The constructive advice from the above-mentioned people has helped me plough through many difficulties and without them this project would have achieved much less.

TABLE OF CONTENTS

ABSTRACT

Protein function prediction has been a key problem in Computational Biology, and traditionally accomplished through “guilt by association” with BLAST. However, when sequence similarity is not available, similarity of other information is utilized to solve the prediction problem. Previous work by Chua *et al* (2007) has developed a software to allow integration of similarity information of various kinds through structuring the protein information into protein pair graph and using protein’s neighbour for majority voting of protein function. In this work, analysis into Pubmed information and simple text mining was performed to better aid protein function prediction under the developed software framework. Organizing of the Pubmed information based on two rules derived from text mining brings about higher precision of function prediction. The Pubmed information graph was organized into disjoint subsets of the Pubmed graph, based on two rules 1) protein pairs occurring in same sentence and 2) abstract contain species-of-interest species name or common name. The organization of Pubmed information graph increased prediction average precision for Gene Ontology (GO) domain Biological Process by **4.2%,** GO domain Cellular Component **8.0%** average rise in precision and **1.7%** average rise in precision for GO domain Molecular Function.

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Protein function prediction has all along been a key problem in computational biology. It has traditionally been accomplished primarily using "guilt by association" of sequence similarity. Basic local alignment search tool (BLAST) is the most popular tool used in this sequence similarity (homology) search (*Altschul et al*, 1990), which works on the assumption of sequence similarity to infer function similarity. However, if good sequence similarity to a previously functionally annotated protein is unavailable, one must appeal to guilt by association of other types of similarity and even to combination of multiple types of similarity information.

Several methods uses similar information of other kinds such as protein-protein interaction (Letovsky and Kasif, 2003), structural features of protein (Arakaki et al, 2004), phylogenetic analysis with integration of experimental and homology information (Sjölander. 2004) and even text mining of literature for protein function prediction with the BioCreative inititive (A critical assessment of text mining methods in molecular biology) (Hirschman *et al*, 2005). With the advancement of technology, wealth of information is available in online databases providing various similarity information, waiting to be tapped into. A recent work has developed a framework for the fusion of multiple types of similarity information to enable effective use of this wealth of information for protein function prediction. Results show that, 1) similarity information fusion method works well, 2) simple co-occurrence count gives reasonable sensitivity & precision, and 3) combining multiple information sources outperforms any single information source (Chua *et al*, 2007). This program is the interest of analysis in this project.

With so many literatures available in online literature database, particularly Pubmed having around 16 million articles in 2007, and still growing. It is a known fact that many knowledge are still locked inside literature text, hence, the BioCreative initiative was began to assess the state of the art for text mining applied to biological problems (Hirschman et al, 2005). BioCreative focused on two tasks. The first deals with extraction of gene name and protein name from text, while the second task addressed issues of functional annotations, requiring system to identify text passage that supported the functional annotation for specific protein. The second task is more relevant to the objective of this project. Projects in BioCreative used various classification methods, including SVM (Mitsumori *et al*, 2005) and Bayesian networks (Ray and Craven, 2005), along with linguistics domain knowledge (Tamames, 2005) to analyse text for single protein annotation. However, we are using the program by Chua *et al* (2007) for this text analysis task, which is based on majority voting of neighbouring protein function for function annotation.

We hypothesize that further analysis and processing of the Pubmed information source for input into the program-of-interest would bring about better performance by the prediction algorithm. We also hope to be able to generate interesting observation which might assist the text mining effort of BioCreative through our simple text analysis here. The analysis of words by chi-square and odds ratio has identified that organism-of-interest species name and common name are useful and that sentence based analysis has a very high probability to infer function sharing for protein pairs occurring in those sentence.

The processing of Pubmed abstract information using organism species name and common name to reorganize the information and further tapping on the high proportion of protein pairs with similar function name in a sentence to further organize the data, has given rise to more precise information than the original Pubmed unprocessed. This organization of data brings about a **10%** increase in precision for the Gene Ontology (GO) domain of Biological Process, on certain region of recall (0.4 to 0.8) and the average rise in precision was about **4.2%** while maintaining the same maximum recall achievable. For GO domain Cellular Component, the result also saw a **10%** rise in precision at certain region of recall and average rise in precision across whole range of recall was **8.0%**. For GO domain Cellular Function the results were not that significantly better with around **1.7%** average rise in precision across the range of recall.

2. MATERIALS AND METHODS

2.1 Materials

In this project, prediction of yeast protein function from Saccharomyces Genome Database (SGD) constitutes the interest of prediction of protein function. The set of 6058 proteins from the database is the object of interest for the prediction of function.

Protein functions to be annotated to the protein are limited to the list of function found in Gene Ontology (GO). GO provides a standardized vocabulary to describe gene and gene product. Gene Ontology provides a hierarchical layout of the function annotation dictionary, with more general annotation at the lower level and more specific annotation at the higher level forming an acyclic graph (Ashburner *et al*, 2000). The use of a

standardized vocabulary is necessary and critical with the large number of scientific group working on the problem, as standardization would allow more efficient information sharing and comparison.

The analysis data objects are the list of Pubmed abstract containing the names and synonyms of the protein from SGD. The abstracts were searched through a querying program, using the assigned identification number from SGD database, along with synonyms of the protein from SGD database, against the Pubmed database. Abstracts inclusion for analysis were only limited to the first 1000 abstracts and an added constraint to limit the search to title and abstracts with the Pubmed filter option. In this project, analysis deals only with abstract and title text; hence this constraint is added to gather data relevant for subsequent analysis.

2.2 Scoring function and Interaction map

In this project, the input to the classification system in use are files of the format having pairs of protein related as inferred from relevant data sources (Pubmed, Pfam, BLAST *etc*), and a score for each pair of proteins. What the input files describe is an **interaction map** between proteins connected by edges with a score for each edge which is used for function prediction in the program. The scoring function used was previously defined by *Chua et al* (2007). However, a pseudo-count of 1 was added here to the denominator to the previous function. This pseudo-count of 1 is a strategy adapted in many previous work (Tibshirani *et al*, 2002 & Tusher *et al*, 2001), and works to prevent giving

large scores to protein pairs with very small set of abstracts. New score function, *S(u,v)*, is formally defined as follow.

$$S(u,v) = \frac{|A_u \cap A_v|}{1+\sqrt{A_u \times A_v}} \quad \textbf{-(1)}$$

A_x is the set of Pubmed abstracts that contain protein x.

This is the major interest of this project; how will different ways and rules pushed into the construction of interaction map affect protein function prediction.

2.3 Prediction of function

Protein function prediction analysis was done with the aid of a protein function prediction tool by *Chua et al* (2007). The program, **predict.pl**, is a *Perl* program and below is a brief introduction to some of the key function and concept of this program.

Based on the scoring function used for the data source (for Pubmed scoring function elaborated in **Section 2.2**), the edges are distributed into twenty baskets of equal score intervals for individual analysis of function transferring. Hence, the scoring function works to distribute the edges into the different basket on assumption that edges with similar score should be similar and contribute towards function prediction positively within the same basket.

Within each basket function is predicted by building an interaction map (or graph) based on the information source (such as Pubmed, BLAST data, Pfam data *etc*). The graph

has nodes, representing proteins, and nodes connected via edges, nodes connected this way by an edge means an implied function sharing between the nodes by the information source and the program predicts function for a protein based on its neighbours' function via weighted voting.

To assign a probability of transferring of function for an edge for one data source, the reliability of the information source is first assessed by a confidence score, which essentially measures the reliability or weight of the information source to suggest function similarity. Function similarity suggested by an information source is affected by a myriad of factors from nature of experiment, noise in experiment down to threshold setting in embedded score of the information source. These factors are summarized by the confidence function, which measures the probability of a data source *k* to transfer function *f*, estimated by:

$$p(k, f) = \frac{\sum_{(u,v)\in E_{kf}} S_f(u,v)}{|E_{kf}| + 1} \quad \textbf{-(1)}$$

where E_{kf} is the subset of edges of data source *k* where each edge has either one or both of its vertices annotated with function *f*.

S(u,v) = 1 if *u* and *v* shares function *f*, 0 otherwise.

As more information sources are provided, the aggregated confidence for the edges over this entire set of information sources is calculated by:

$$r_{u,v,f} = 1 - \prod_{k\in D_{u,v}} (1 - p(k, f)) \quad \textbf{-(2)}$$

$D_{u,v}$ is the set of data sources which contain edge *(u,v)*.

Lastly, assigning score of an annotation to a protein in the completed map built from the various information sources is defined by:

$$S_f(u) = \frac{\sum_{v \in N_u} (e_f(v) \times r_{u,v,f})}{1 + \sum_{v \in N_u} r_{u,v,f}} \quad \textbf{-(3)}$$

$S_f(u)$ is the score of function *f* for protein *u*,

$E_f(v)$ = 1 if protein *v* has function *f*, 0 otherwise,

Nu is the set of neighbours of protein *u*,

$r_{u,v,f}$ is the link confidence between proteins *u* and protein *v*.

This program is found to work well, with reasonable precision-recall level and efficient for projection and inclusion of multiple data sets as stated in the paper by *Chua et al* (2007).

2.4 Assessment of prediction performance

In this project, the aim is to extract some rules from Pubmed abstracts to aid prediction of protein function. The software which is used for function prediction is the software, **predict.pl,** described briefly in **Section 2.2**. The program outputs information with respect to its performance in cross-validation mode. Output includes the

1) **Number of annotated terms predicted** at various **Receiver Operating Characteristics (ROC)** score and

2) **Recall** and **Precision** of the prediction made at various ROC score thresholds.

ROC graph is a plot of true positive rate versus false positive rate, with a list of classifiers at different thresholds of $S_f(u)$ in **Equation 3**, for assigning protein function. The plot of the different classifiers at different thresholds on the ROC gives rise to a graph and the area under the curve (ROC score) is a popular measure of the discriminative ability of the classifier. The machine learning community most often uses the ROC AUC statistic. This measure can be interpreted as the probability that when we randomly pick one positive and one negative example, the classifier will assign a higher score to the positive example than to the negative. However, with so many terms to consider, the output is summarized by plotting the number of informative GO terms that can be predicted with ROC scores better than or equal to various ROC thresholds. False positive (*fp*) rate and true positive (*tp*) rate are calculated as:

$$fp = \frac{FP}{N} \qquad \textbf{-(4)} \qquad\qquad tp = \frac{TP}{P} \qquad \textbf{-(5)}$$

FP is false positive,

TP is true positive,

N is negative class,

P is positive class in the contingency matrix of a classifier.

According to Chua *et al*, 2007 and Fawcett, 2004, ROC score captures the discriminative power of the classifier to assign a function to a protein but ROC score does not capture how accurate the classifier is, to capture that performance requires the precision and recall of the classifier.

$$\Pr ecision = \frac{\sum_i^K k_i}{\sum_i^K m_i} \qquad \textbf{-(6)} \qquad \Re call = \frac{\sum_i^K k_i}{\sum_i^K n_i} \qquad \textbf{-(7)}$$

k_i is the number of functions correctly predicted for protein i;

m_i is the number of functions predicted for protein i; and

n_i is the number of functions annotated for protein i.

With the precision and recall, a summarizing figure for the graph of precision versus recall was introduced, named the Area Under the Precision Recall Curve (**AUPRC**). The AUPRC is an estimation of the area under the precision versus recall curve, and the difference of AUPRC between two systems give an indication of the average rise or drop in precision across the range of recall when comparison is needed. This is calculated using a sixth power polynomial equation generated by Microsoft Excel to estimate the precision versus recall graph, area under the curve is determined via integrating this function. The graphs generated via Microsoft Excel and equations are presented in **APPENDIX B** for reference.

With the 2 main performance indicators and a derived one, we are able to tell how good is the classifier at separating the function of a protein from functions it does not have, and further with precision and recall we can tell how often this score is correct. The

program's main advantage is its ability to take multiple sources of data into consideration, which in previous work (*Chua et al*, 2007) has been confirmed that taking more data sources outperform any single data source alone. Hence, in line with the observation above, we would also want our analysis of Pubmed to be able to translate to better precision even in combination with other data sources such as BLAST or PFAM, versus unprocessed Pubmed abstract interaction map with other data source.

2.5 Chi-square test and Odds ratio analysis

Pubmed text being the data object in this project requires the use of text handling methods to enable its analysis. Firstly, Stop words, or stopwords, are removed from the text string from further processing. Stop words is the name given to words which are filtered out prior to, or after, processing of natural language data (text), which include "to", "that" and "and". Next, the remaining words are "stemmed" to the root form using Porter's Stem, example of stemming would be from "consideration" to "consider".

In the course of the project, analysis on the dependence or independence of words to appear in abstracts which suggest a function similarity between a pair of protein which appears in the same abstract was done. To do this we used the chi-square statistical test to test the null hypothesis that words have no preference for either of 2 text types, **positive abstract** and **negative abstract**. In this case abstract with at least a pair of protein sharing one GO function name are labeled **positive abstract**; and abstract with all protein pairs not sharing any function name are labeled **negative abstract**. To remove words which do not appear frequent enough in the set of abstract we set a minimum threshold appearance of

above **10%** appearance in total number of abstract. This is to remove words which are infrequent in abstract and to prevent redundant analysis on words which are not frequent enough to be projected into general cases. Next chi-square significant words at chi-square score of greater than **3.84** ($p \leq 0.05$) were further analyzed via log odds ratio to detect words which are significantly expressed in positive abstracts.

Chi-square test was done using the formula below, with the setup of a 2-by-2 contingency table.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad \textbf{-(8)}$$

$i \in \{word, others\}$
$j \in \{positive, negative\}$

$O_{i,j}$ = *Observed frequency of i and j*
$E_{i,j}$ = *Expected frequency of i and j*

Table 1. shows an example of the 2-by-2 contingency table (from data) and following that the chi-square calculation for the example and expected value is calculated from the observed contingency table as follow.

Expected value = (row total * column total)/ grand total

Calculation for Expected value for positive abstract containing “yeast” keyword:

Expected value = [(g) / (i)] * (e)

= 15600/192263 * 27221

= 2208.68

	Observed counts			Expected values		
	Positive	Negative	Total	Positive	Negative	Total
Yeast	5908 **(a)**	21313 **(b)**	27221 **(e)**	2208.68	25012.32	27221
No yeast	9692 **(c)**	155350 **(d)**	165042 **(f)**	13391.32	151650.78	165042
Total	15600 **(g)**	176663 **(h)**	192263 **(i)**	15600	176663	192263

Table 1. Example of contingency table for chi-square analysis on the word "yeast"

;

Chi-square calculation for "yeast":

$$\chi^2 = \frac{(5908-2208.68)^2}{2208.68} + \frac{(21313-25012.32)^2}{25012.32} + \frac{(9692-13391.32)^2}{13391.32} + \frac{(155350-151650.78)^2}{151650.78}$$
$$= 6195.99 + 547.13 + 1021.93 + 90.24$$
$$= 7855.29$$

In addition to significantly biased words to appear in either text type, the identification of which direction the word is more biased towards, in this case finding words biased to appear in positive abstract, is needed. To accomplish that, the commonly used log odds ratio is used, as defined:

$$\textbf{log Odds Ratio score} = \log_{10}\left(\frac{a*d}{b*c}\right) \qquad \textbf{-(9)}$$

Taking "yeast" keyword as an example again,

Log Odds Ratio score = $\log\frac{5908\times155350}{21313\times9692}$ = 0.648 (3 s.f.).

2.6 Sentence based Interaction map

A hypothesis that protein occurs in the same sentence are more likely to share function than proteins which occur in the same abstract but in different sentence is put up. This hypothesis stems from the general understanding of how individuals construct

sentence, especially in the case of biological studies report (Pubmed abstract). Scientist would want to make some inference of the function or link proteins together in their studies; to achieve that usually would involve inclusion of these proteins in a single sentence. Hence, this hypothesis went under scrutiny in the course of this project, whereby only proteins of prediction interest co-occurring in the same sentence in Pubmed abstract was included in the interaction map construction.

2.7 Segmenting versus Filtering

Filtering of edges would give rise to better precision with irrelevant edges being discarded, however, this advantage comes at a price of reduced discriminative power of the classifier, since filtering removes edges from consideration, the graph built up has less and less edges hence contributing to the reduced discriminative power of the classifier. There are two approaches to solve this problem, **1)** the inclusion of the full population of edges along with the precise interaction map built from filtering as described in **Section 2.6**, or **2)** segmenting the total population of edges instead of filtering them. The latter approach is preferred in this case as discussed in **Section 3.3**.

The total population of edges from Pubmed abstract was segmented into different sub-data source, instead of filtering the edges. With this approach, segmentation of the total edges from Pubmed abstract into three sub-data sources was done and fed into the program for prediction, the 3 sub-data source are segmented accordingly as listed

1) Protein pairs co-occurring in abstracts with genus name and co-occurring in the same sentence in those abstracts

2) Protein pairs in abstracts with genus name but protein are not mentioned in same sentence,

3) Protein pairs occurring in abstracts which did not mention genus name at all.

2.8 Word analysis of sentence containing protein name

Pubmed abstract were analyzed on the sentence level to screen for words which might suggest transfer of function between a pair of proteins in the sentence when occurring together. Pubmed abstracts were first broken down into sentences, next protein name were search for in the sentences, lastly only those sentences with protein names were parsed to text handling and chi-square analysis as described in **Section 2.5**.

Next, the edges in **Section 2.7** were further segmented into the following

1) Protein pairs co-occurring in abstracts with genus name and co-occurring in the same sentence in those abstracts with the keywords found

2) Protein pairs co-occurring in abstracts with genus name and co-occurring in the same sentence in those abstracts without the keywords in the sentence

3) Protein pairs in abstracts with genus name but protein are not mentioned in same sentence,

4) Protein pairs occurring in abstracts which did not mention genus name at all.

3. RESULTS

3.1 Chi-square analysis of words in abstract

The list of chi-square significant words (**chi-square score > 3.84**) is attached in **Appendix A**, presented in **Table 2** is the top ten words of **Appendix A**. The words are arranged according to their log odds ratio score, we found that the organism species name and common name is ranked the top three words (log odds ratio over 0.6 implying over four times more likely to appear in positive abstract) which are associated to positive abstracts, words such as “yeast”, “saccharomyces” and “cerevisiae”. In the rest of the report we shall refer to the 3 word referring to the yeast, including “yeast”, “saccharomyces” and “cerevisiae”, simply as “**genus**”. This propelled us to do a preliminary filtering of the abstract based on the genus keyword and to do a co-occurrence map based on this set of filtered abstracts containing the genus name. **Figure 1** and **2** presents the results based on the preliminary results we get, namely the accuracy or reliability of the classification (**Figure 1**) and the discriminative power of the classifier at various ROC score threshold (**Figure 2**). “**Genus filtered**” is the graph for the result of prediction done with the interaction map of protein pairs in abstract containing the genus name (filtering).

In **Figure 1** and **2**, the interaction map built from unprocessed Pubmed was included as “basal reference”. In **Figure 1**, the “Genus filtered” graph had an improvement of **5%** precision at recall level of 0.4 to 0.8 compared to the basal reference; while at high level of recall (>0.8) precision converges rapidly. The AUPRC is presented in **Table 3**, with “Genus filtered” having a minor increase in AUPRC of **0.022** (average 2.2% increase in precision).

From **Figure 2**, the discriminative power of the classifier is maintain with a mild drop of 1% (1 less predicted term out of 105 annotated terms) at ROC score of 0.9, which we find acceptable, this drop is expected since when we filter out abstracts with the genus keywords, we are ultimately removing edges from consideration in the final graph, this would contribute to a slight drop in discriminative power since less edge information is in the graph now.

		Observed value					
word	Total	Positive (A)	Negative (B)	Not_word positive (C)	Not_word negative (D)	chi	OR
cerevisia	18637	4836	13801	10764	162862	8804.403	0.724422
saccharomyc	18055	4684	13371	10916	163292	8495.786	0.719354
yeast	27221	5908	21313	9692	155350	7855.286	0.647694
defect	13697	2621	11076	12979	165587	2402.925	0.479869
mutant	31411	5159	26252	10441	150411	3477.806	0.451941
requir	30590	4662	25928	10938	150735	2477.981	0.394079
homolog	12089	2003	10086	13597	166577	1236.892	0.386134
subunit	13899	2244	11655	13356	165008	1296.134	0.376339
complex	30757	4564	26193	11036	150470	2221.051	0.375799
deletion	11830	1914	9916	13686	166747	1099.838	0.371387

Table 2. List of top10 log odds ratio significant words which are chi-significant

3.2 Prediction of sentence based interaction map

Based on the hypothesis in **Section 2.6**, a pre-analysis of an abstract's ratio of correct edges to total edges inferable from a data source was done and presented the analysis as a bar chart of the total number of abstract with that ratio of correct edges versus ratio of correct edges. Furthermore, the ratio of correct edges was further subdivided to show the number of proteins in the abstract which reflects the number of edges inferable.

Figure 3 presents the abstract level information while **Figure 4** presents the sentence level information for the GO domain Biological Process, the result for the other 2 domain are included in **APPENDIX C** for reference. From the data drawn to construct **Figure 3**, **52%** (2390 out of 4580) of the abstract contains less than 50% edges which have both proteins sharing a function. Data used for construction of the two figures (inclusive of the Cellular Component and Molecular Function) are included in **APPENDIX D**. Bearing in mind that with each added protein mentioned in an abstract our number of edges inferable raises quadratically according to the formula defined:

$$edges = \frac{(prot\_no) \times (prot\_no - 1)}{2} \quad \textbf{-(11)}$$

where *prot_no* is number of protein in abstract.

This means that when using abstract data as a whole, we are including all the possible edges inferable from the protein pairs existing in the abstract, inevitably we are including other edges in the abstract which are not having the protein pair sharing function. Hence, we would like to avoid that by drawing an edge only when the protein is mentioned in the same sentence.

Based on this hypothesis, an interaction map was constructed based on protein pairs that occurred in the same sentence only, and with this map the results are presented in **Figure 1** and **2,** labeled as **"Genus sentence level"**. Based on the graph for Biological Process GO domain, there is significant improvement of precision of as much as **10%**

compared to basal reference at the respective recall level from 0.4 to 0.8, this improvement was also evident in Cellular Component GO domain; but was not significant in Molecular Function. AUPRC scores are also listed in **Table 3**. However, upon closer inspection of the discriminative power of the classifier (Annotated terms predicted), there is a **10%** (10 less predicted terms out of 105) reduction in the number of predicted annotated terms at ROC score of 0.8 to 0.9. This reduction of discriminative power is quite significant and we would like to ensure that discriminative power of the classifier does not suffer in the processing of the data.

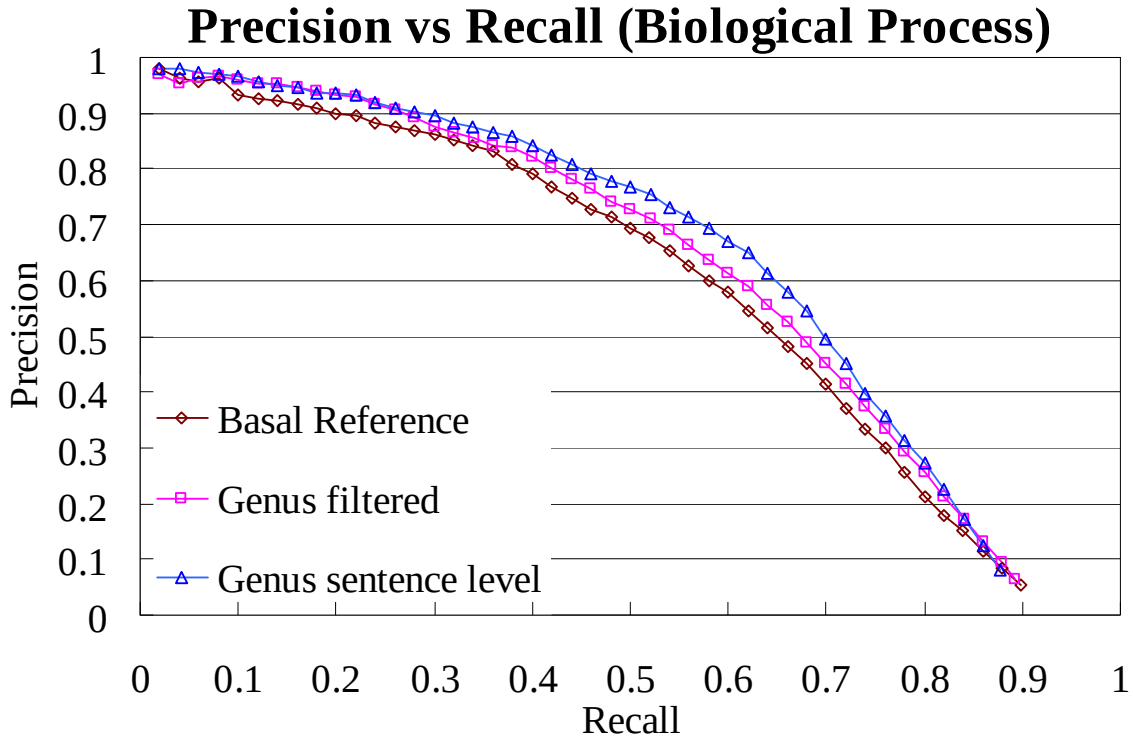


Figure 1a. Precision vs. Recall graph: Biological Process

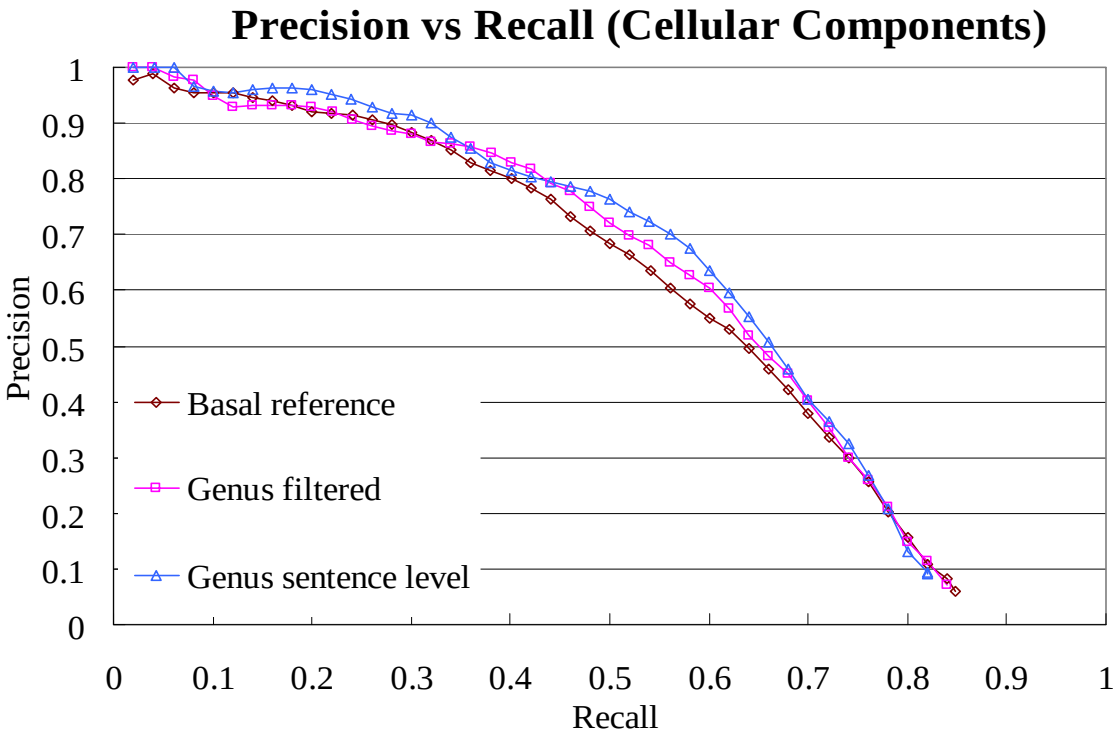


Figure 1b. Precision vs. Recall graph: Cellular Component

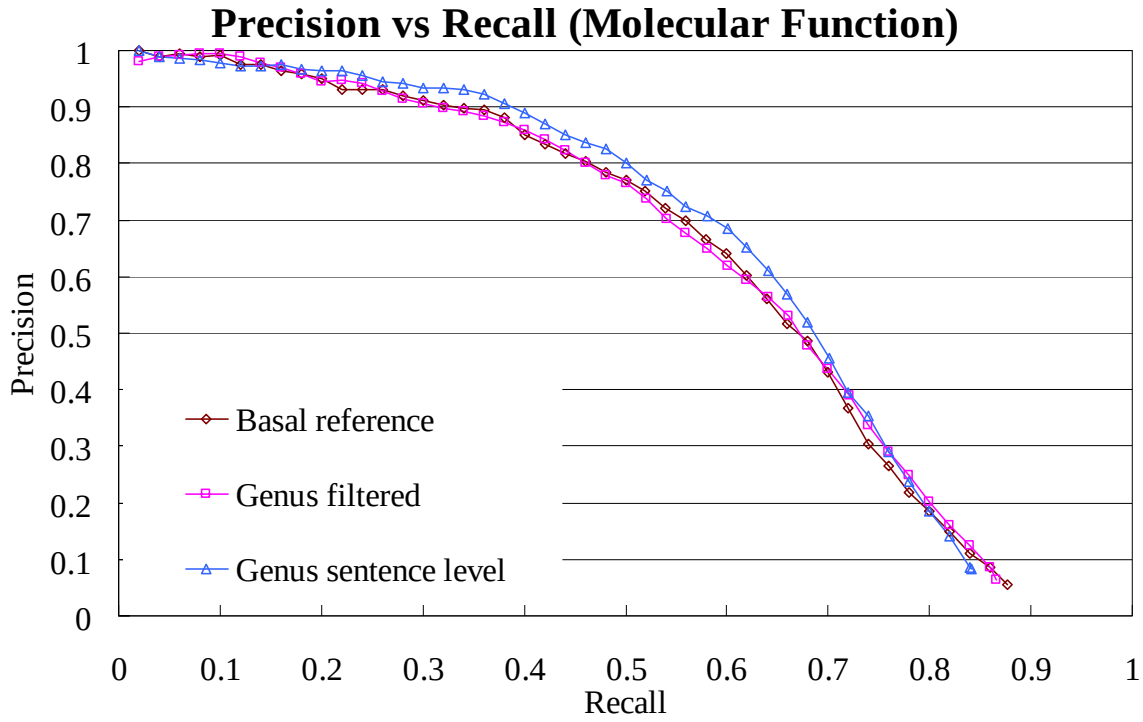


Figure 1c. Precision vs. Recall graph: Molecular Function

Figure 1 presents the precision versus recall graph of the function prediction for "Basal reference", "Genus filtered" and "Genus sentence level". **a)** GO domain: Biological Process **b)** GO domain: Cellular Component **c)** GO domain: Molecular Function.

Interaction map	**Domain AUPRC score**		
	Biological Process	**Cellular Component**	**Molecular Function**
Basal reference	0.595	0.576	0.643
Genus Filtered	0.617	0.588	0.619
Genus sentence level	0.640	0.646	0.654

Table 3. AUPRC score of 3 domain for "basal reference", "Genus filtered" and "Genus sentence level"

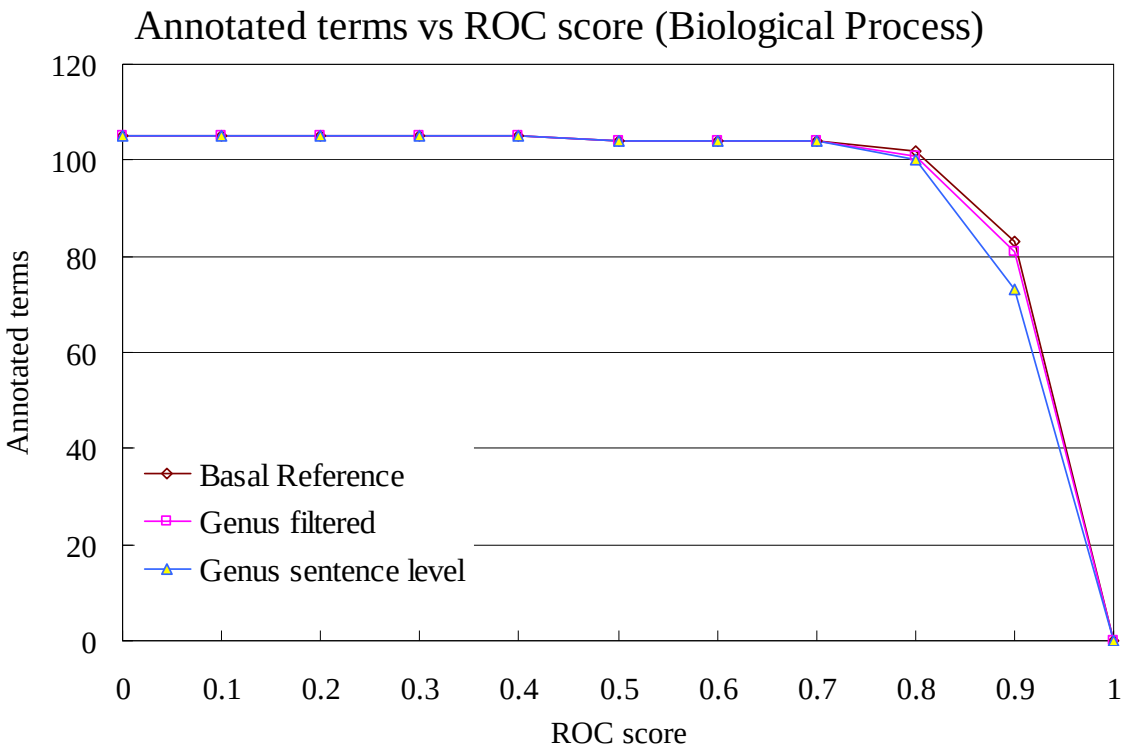


Figure 2a. Annotated Terms vs ROC score: Biological Process

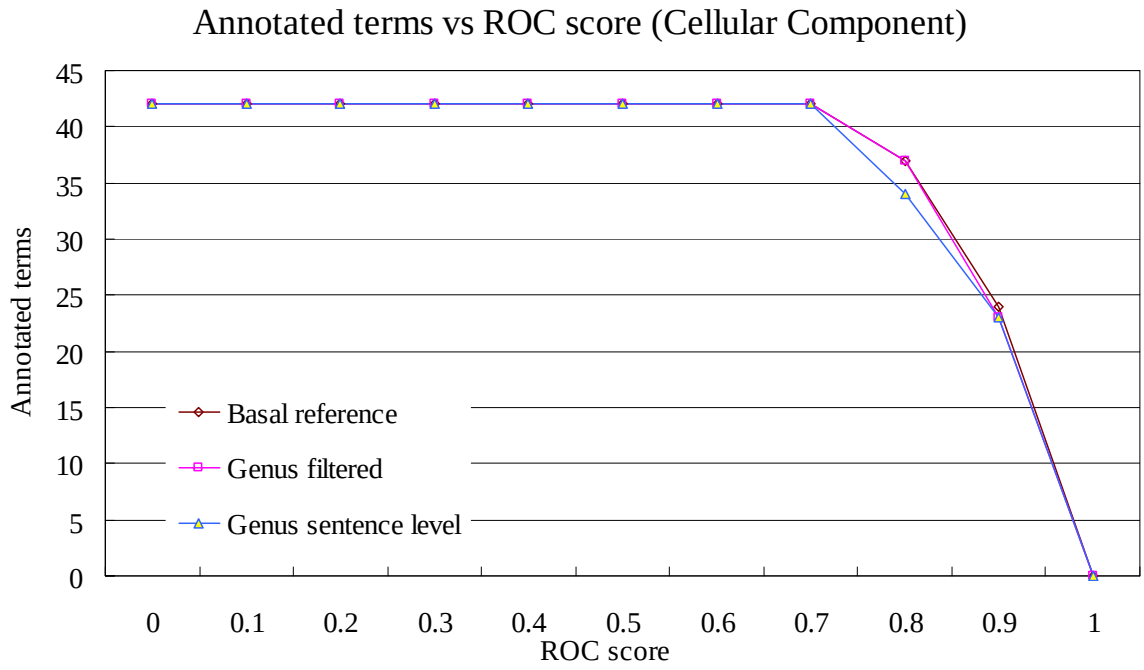


Figure 2b. Annotated Terms vs ROC score: Cellular Component

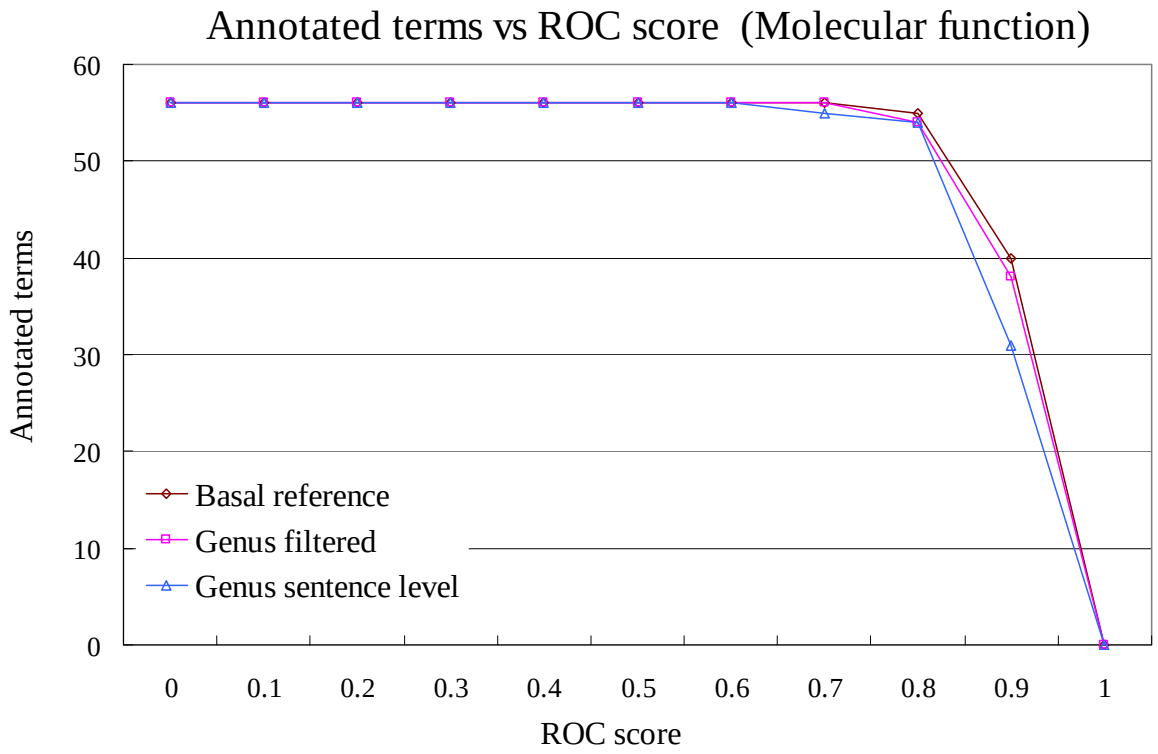


Figure 2c. Annotated Terms vs ROC score: Molecular Function

Figure 2. presents the annotated term versus ROC score of "Basal reference", "Genus filtered" and "Genus sentence level". **a)** GO domain: Biological Process **b)** GO domain: Cellular Component **c)** GO domain: Molecular Function.

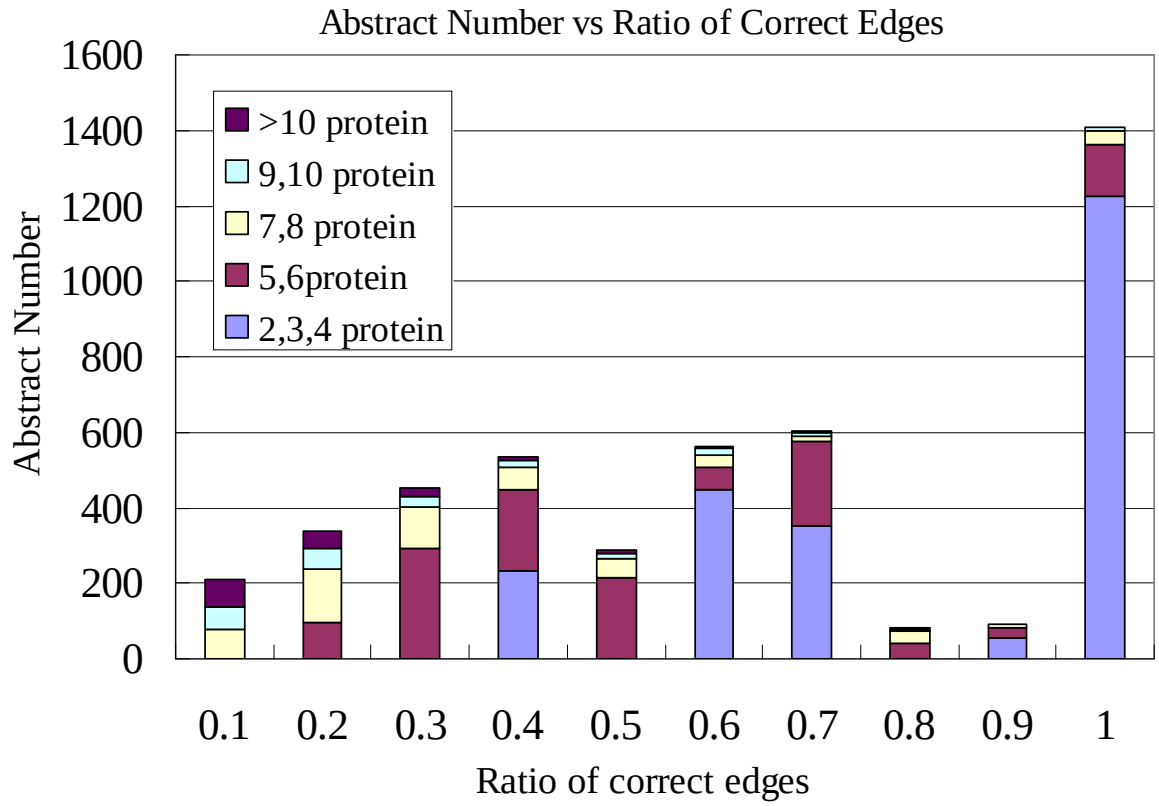


Figure 3. Abstract number versus ratio of correct edges in Domain: Biological Process

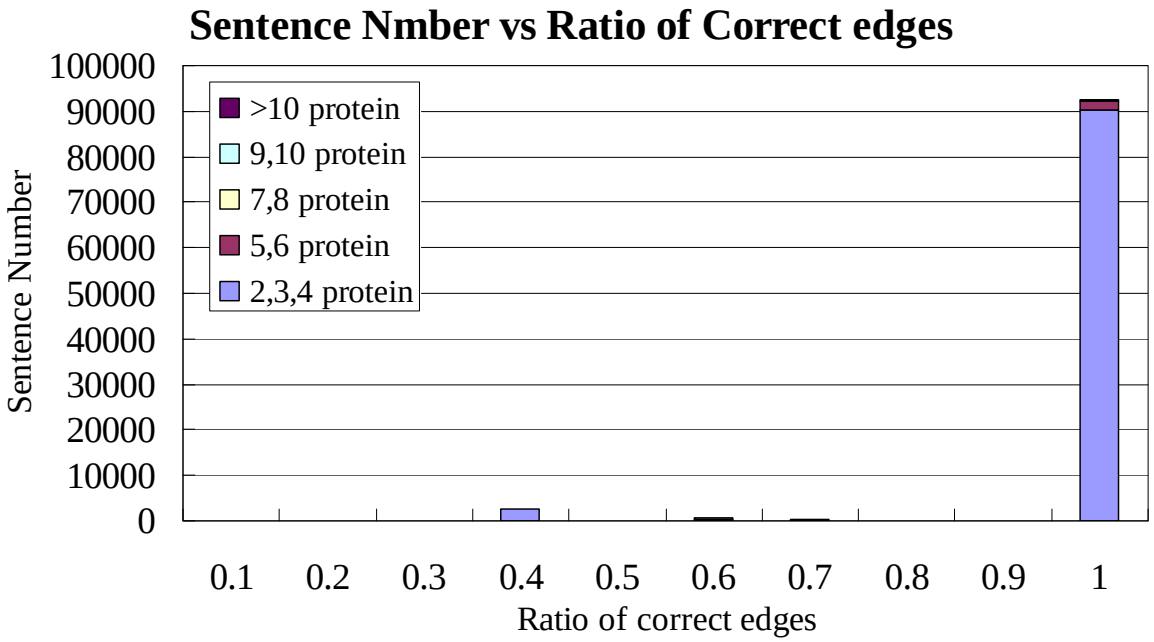


Figure 4. Sentence number versus ratio of correct edges in Domain: Biological Process

Comparing **Figure 3** and **4**, there is a very stout difference between the two charts, with the sentence level information having more protein pairs with high ratio of correct edges compared to abstract. Around **52%** of abstract have less than 50% of correct edges with respect to the total number of edges possible, while over 90% of sentences have all protein pairs sharing a function name. Furthermore, the positive pairs occur mostly in sentences with 2 or 3 proteins mentioned in them. Hence, sentence based analysis gives better precision than abstract based.

3.3 Segmenting of interaction map

From **Section 3.2**, there was an improvement in the precision of the classifier with filtering of edges, however, as **Figure 2** shows, for the three domains the discriminative power of the classifier dropped at certain ROC score threshold (0.8 to 0.9), hence, a new approach was taken to address this problem. In this approach filtering of the edges was not done; instead the total number of edges was segmented into sub-sources. Presented in **Figure 5** is the classification result of the segmenting of the total population of edges inferred from Pubmed abstract into the three sub-sources (as laid out in **Section 2.7**) labeled "**Segmented**" compared against the classification result of the combination of two sources (labeled "**2 sources**"), 1) the interaction map from **Section 3.2** (inclusion of protein pairs co-occurring in sentence and in abstract containing genus name) and 2) the interaction map built from the total abstract (no processing). Unprocessed Pubmed interaction map prediction result is included as "Basal reference". **Table 4**. presents the AUPRC of the different graphs.

With the results from **Figure 5** and **Table 4**, for GO domain Biological Process, we can see that segmenting of total edges gives better precision of up to **5%** at the recall level of 0.4 to 0.8 and yet discriminative power is not reduced, furthermore, AUPRC score is **2.5%** higher than "**2 sources**", meaning that the averaged raise of precision in the precision versus recall graph is about **2.5%** given that both graph has the same maximum recall. The results are similar for the other 2 domain. **Figure 6** show that the processing did not affect discriminative power at all.

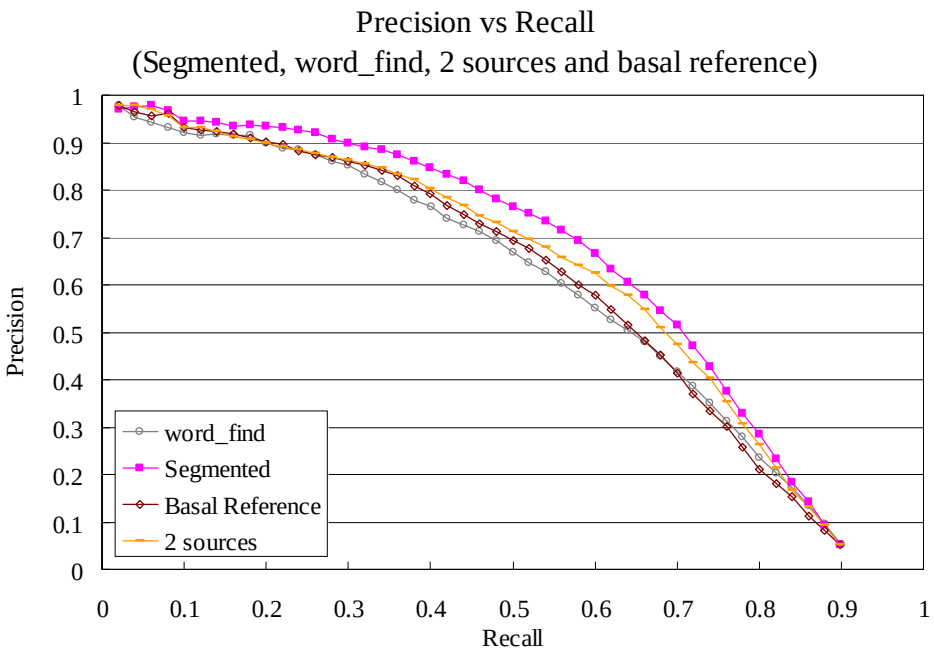


Figure 5a. Precision vs Recall: **Biological Process**

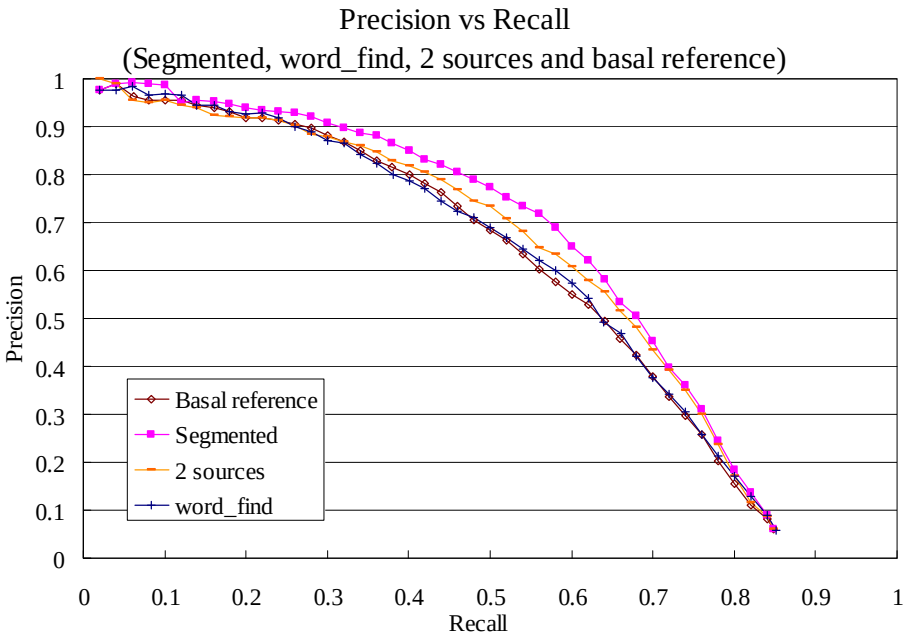


Figure 5b. Precision vs Recall: **Cellular Component**

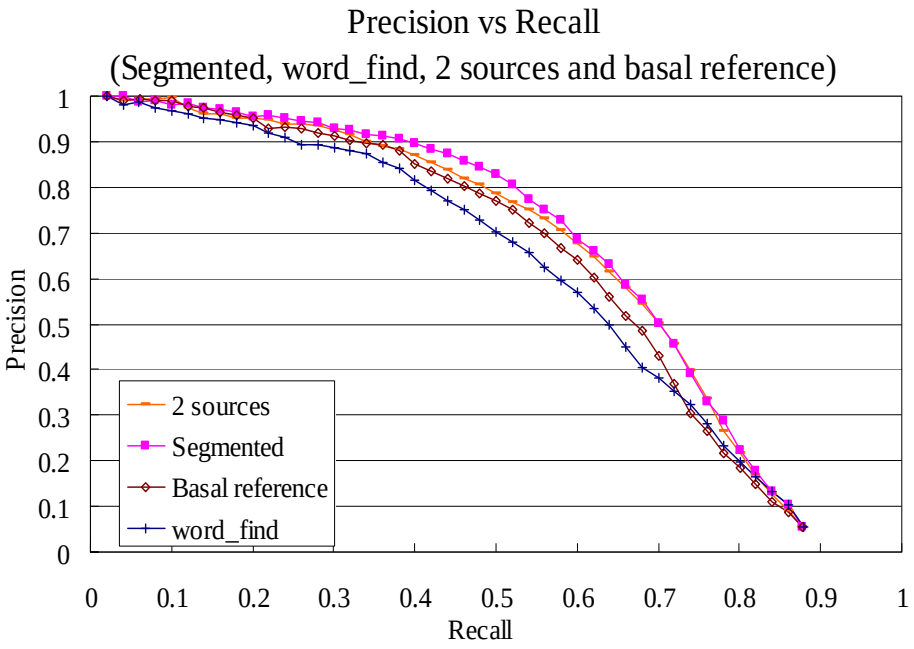


Figure 5c. Precision vs Recall: **Molecular Function**

Figure 5. present the Precision versus Recall of "Basal reference", "2 sources", "Segmented" and "word_find". **a)** GO domain: Biological Process **b)** GO domain: Cellular Component **c)** GO domain: Molecular Function.

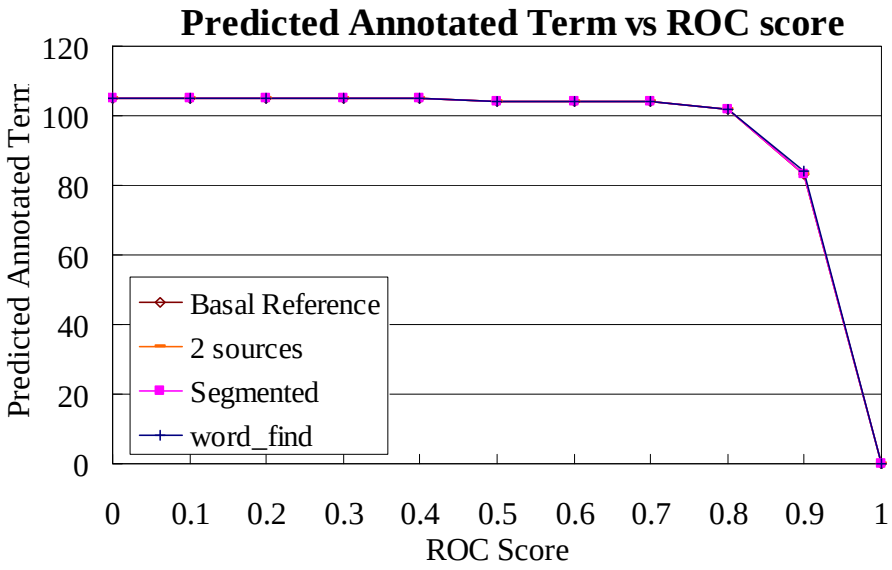


Figure 6a. Predicted Annotated term vs ROC: **Biological Process**

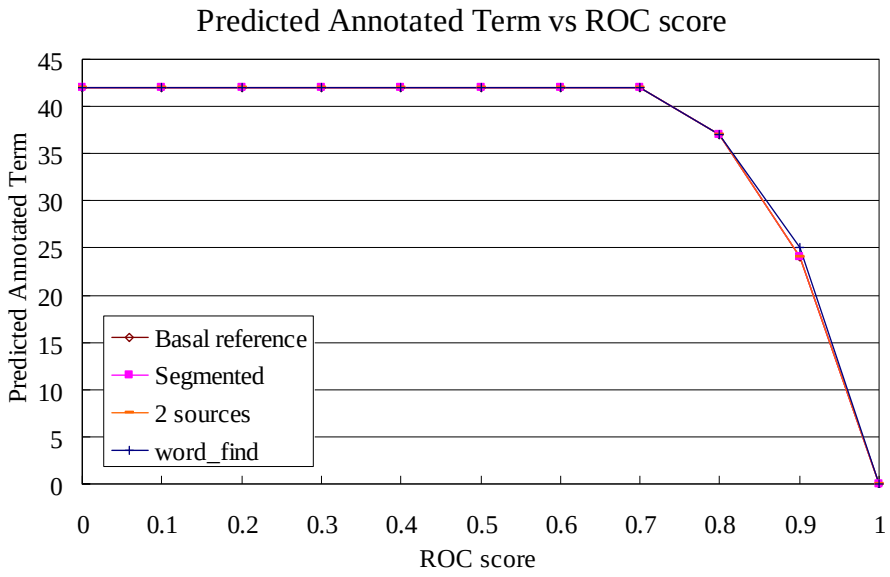


Figure 6b. Predicted Annotated term vs ROC: **Cellular Component**

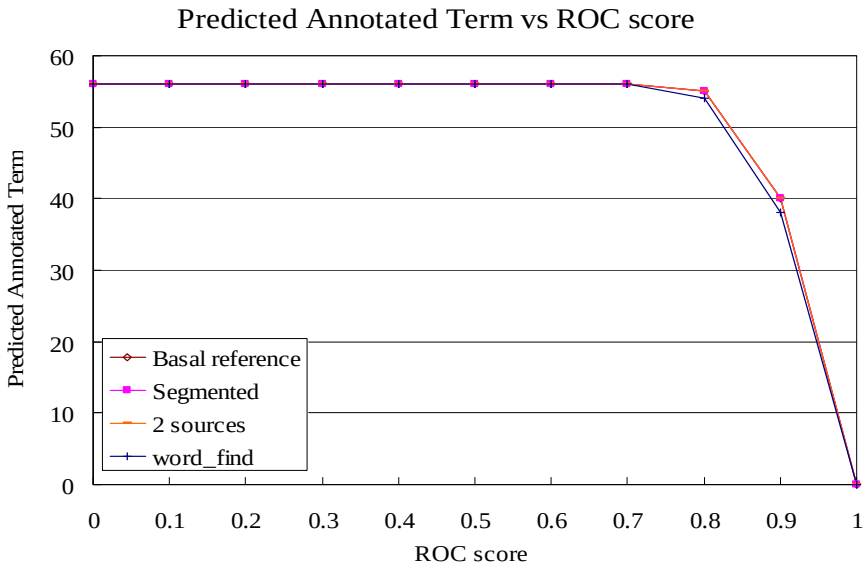


Figure 6c. Predicted Annotated term vs ROC: **Molecular Function**

Figure 6. present the annotated term versus ROC score of "Basal reference", "2 sources", "Segmented" and "word_find". **a)** GO domain: Biological Process **b)** GO domain: Cellular Component **c)** GO domain: Molecular Function.

	Biological Process	Cellular Component	Molecular Function
Basal Reference	0.595	0.576	0.643
Word_find	0.585	0.590	0.591
2 Sources	0.612	0.590	0.653
Segmented	0.637	0.613	0.660

Table 4. AUPRC score of "Basal reference", "Word_find", "2 Sources" and "Segmented" for 3 GO domain.

3.4 Keyword analysis

Further analysis of the words in the sentences as described in **Section 2.8**, produced the following chi-significant words for the three GO domains, "gene", "active", "express", "protein" and "cell". Further, we included some words which might be useful to suggest transfer of function into the search list including words like "mutant", "bind" and "complex". The prediction results of the sentence containing the keywords we have identified is presented in **Figure 5** and **6**, labeled as "**word_find**". The word searching analysis seems to perform equally or worse than the basal reference for all 3 GO domains.

3.5 Combination with other data sources

The function prediction system employed in this project is capable and directed towards integration of numerous data source for protein function prediction. Hence, with the result from Pubmed analysis ("Segmented" interaction map), we would like to compare the difference in performance between the unprocessed Pubmed abstract interaction map with another data source versus segmented Pubmed interaction map with the same data source. We would base our analysis on the commonly used BLAST data and Pfam data as a representation. **Figure 7** shows the result for "Basal reference", BLAST data ("GOBLAST"), "Segmented", Pubmed unprocessed with BLAST ("Basal reference +

GOBLAST") and segmented with BLAST ("Segmented + GOBLAST"), and **Figure 8** respectively for Pfam with Pubmed.

The prediction performance for "Segmented" was better than the combination of unprocessed Pubmed with Pfam, and combination of "Segmented" with either BLAST or Pfam was better compared to either information source combined with unprocessed Pubmed.

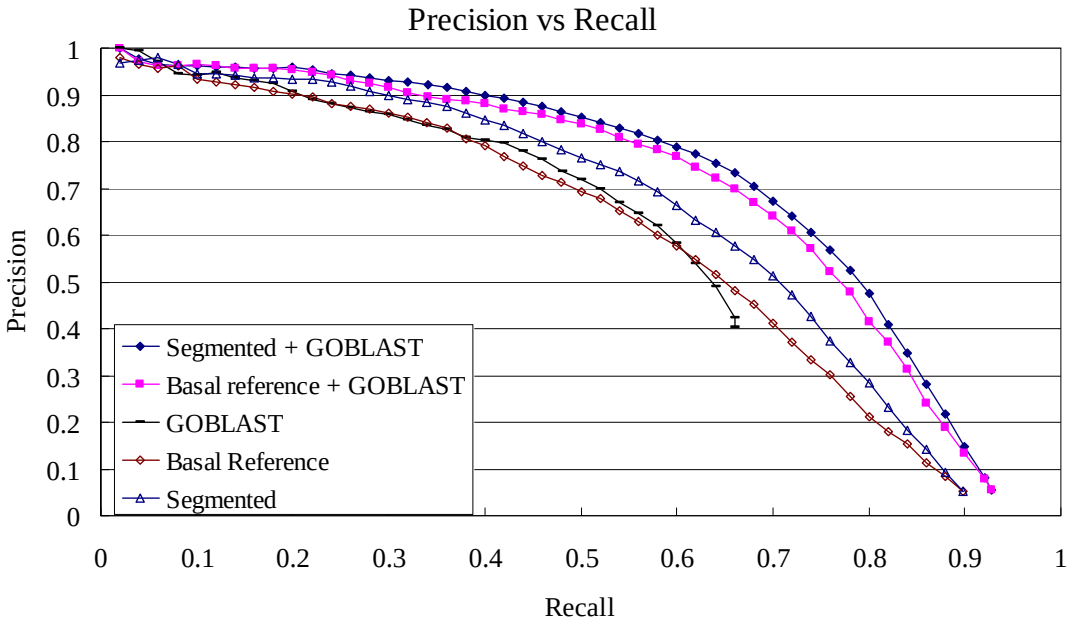


Figure 7. Precision vs Recall of "Basal reference", "GOBLAST", "Segmented", "Basal reference + GOBLAST" and "Segmented + GOBLAST".

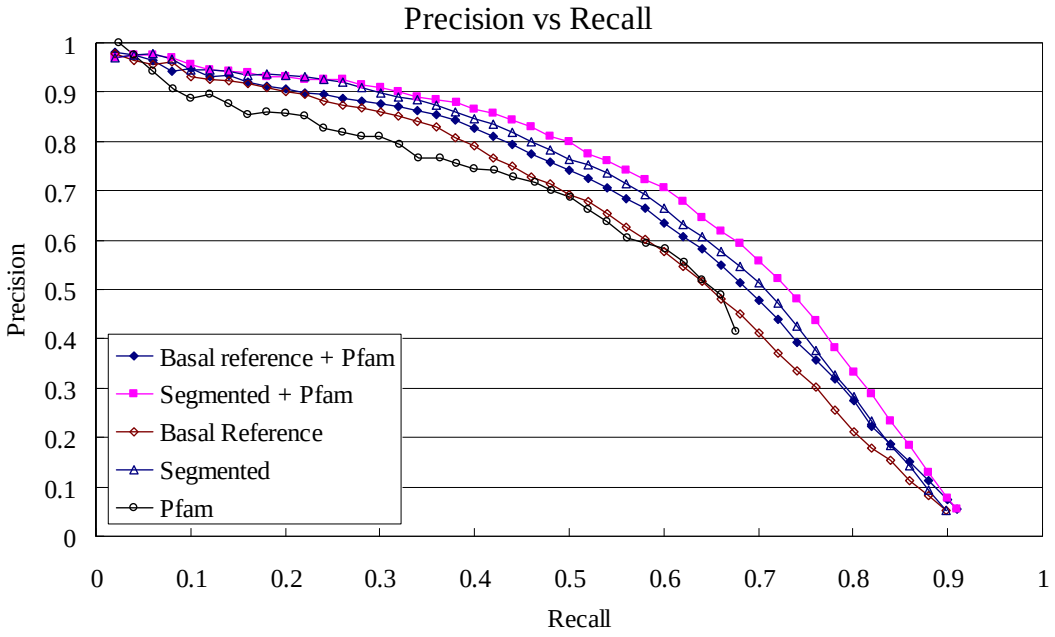


Figure 8. Precision vs Recall of "Basal reference", "Pfam", "Segmented", "Basal reference + Pfam" and "Segmented + Pfam".

4. DISCUSSION

Species name and common name of organism contributes a layer of precise information to the abstract handling. Abstracts are included for interaction map construction based on presence of the protein name in the abstract. Information from homolog protein of other species might be included which might differ from the species we are interested in. In this framework of function prediction, genus keyword suggests a high likelihood of pairs of protein sharing function name from the odds ratio score within the same abstract and was shown to help in function prediction.

Pubmed abstract analysis was based on a simple assignment of whether the abstract transfers a function to the protein based on whether a pair of protein sharing a function

exist in that same abstract. However, the abstract might not be containing just 2 proteins. There might be several other protein pairs present in the same abstract and contributing negative information and considered positive just on the basis of co-occurring with one protein pair sharing function. Therefore, abstract-level analysis for the construction of the interaction map is noisy given that more than 2 proteins are usually mentioned in the same abstract and there might be a lot of irrelevant protein pairs which are included into the map as shown by **Figure 3**. Based on these observations, labeling of articles as positive (transferring of function) on abstract level is not useful in our framework of classification, and this concept is also employed on the general text mining classification problems; whereby usually text in close proximity of the protein of interest is analysed in BioCreative related work, usually within the same sentence (Couto *et al*, 2005 and Verspoor *et al*, 2005). Simple analysis of sentence level information also show that protein pairs are more likely to share some function within the same sentence in **Figure 4**. Interaction map built from protein pairs occurring in same sentence occurring in abstract containing genus keyword was proven to aid in improved prediction as shown in **Figure 1** and **7**, which coincides with the methodology of BioCreative text analysis.

Looking at **Figure 3** and **4**, we might be asking why when the correct edges are so predominant in sentence co-occurrences and yet did not translate to a very sharp rise in precision for the protein function prediction problem. This is due to the simple labeling of an edge as correct (or positive) when the connected protein pairs have a similar GO annotated function name, this does not translate to mean that the protein pair has their list of function name equal to each other, but only at least one of their function name are the

same. This also explains why despite having so many positive protein pairs on sentence level, precision did not rise sharply. Positive pair only contribute positively to the function which the protein pair share function, but for other function which they do not share the protein pair contributes to the denominator of the confidence function (**Equation 1**). Hence, this strategy of labeling correct edges is quite simple and achieves the identification of protein pairs with overlapping function name. Our objective is also satisfied which is to identify the context whereby these positive edges might appear with high chance and at the same time removing those edges with protein pairs that do not share function from consideration with the positive edges.

From the chi-square significant words on abstract level and sentence level, some of the words which are very chi significant are not very high scoring in terms of odds ratio, words including “bind”, “interact” and “express”, which in natural language context suggest relationship between proteins. Pubmed abstracts are summary of scientific finding or experimentally verified relationship between proteins of interest in those papers, where the usual interest is in reporting positive relationship and seldom reporting a negative relationship, unless the negative relationship is interesting (rare). Hence in terms of word usage there should not be too much of a difference using statistical analysis.

Word analysis on the sentence containing protein names was not necessary in the framework of the prediction tool used. In this software, function annotation was determined by majority voting from the protein’s neighbours as elaborated in **Equation 1**. Hence, as **Figure 4** shows sentence provides an already very concentrated positive pairs of protein.

The advantage of this method is the redundancy of intensive text and natural language analysis as done in BioCreative, whereby function annotation is derived from the protein's neighbouring text with a dictionary of the function with some machine learning or voting function to predict annotation. Word analysis in our case only serve to pick up words which might suggest with high chance a sharing of function between the pair of protein in the sentence; essentially words are used to infer high quality pair of protein and what is important is still the protein's neighbouring protein. Disadvantage of this approach is also highlighted here that protein in isolation in the text will not be able to participate in the prediction since it cannot form paired entity in the interaction map, hence a wide variety of information source is needed in the hope that the wide use of information sources would cover each other isolated protein cases.

Next question is, why segment and not filter? As elaborated in **Section 3.3**, the result we presented convinced us but what is the framework or underlying idea that supports this is our next interest. "**Segmented**" interaction map is made up of Pubmed interaction map segmented into sub-sources which are disjoint while "**2 sources**" is a combination of "**Genus sentence level**", the filtered interaction map, and the whole Pubmed interaction map unprocessed. Thus the interaction map "**Genus sentence level**" is a subset of the information from Pubmed. This combination gives that part of the interaction map an addition vote in the final map, which is giving the final decision making equation a high dependence on that part of the map. This dependence makes the decision making equation, **Equation 3**, more unlikely to overturn some of the function assignment decision which contributed to lower precision. Disjoint interaction map performed better

than interaction map with overlapping edges. Hence, independent disjoint interaction map is favourable for prediction in this framework; dependencies in the interaction map used would lower precision, and this justifies the segmentation of interaction map into disjoint sets.

We notice from **Figure 7** and **8** that both Pfam and BLAST data source on their own performed poorly in terms of maximum recall achievable but combined with Pubmed was able to achieve better recall as well as precision and the performance was even better when combined with processed Pubmed ("Segmented"). Furthermore, in **Figure 8** there was a surprise that "Segmented" outperform the combination of Pubmed (unprocessed) with Pfam. This analysis highlighted the richness of information in Pubmed being able to provide high recall level compared to the other 2 information source. Furthermore, the proper processing of Pubmed could even make some information source redundant.

Lastly, from **Figure 1** and **7**, we can notice that processing of Pubmed data did not help to increase recall significantly but only raised precision. The organization of Pubmed interaction map ultimately gives us more precise information which allows the prediction to reduce the number of false positive prediction which contributed to higher precision, however, rising recall requires more than just precise information. To raise recall would require the prediction of more function (less false negative); this would depend on the amount of information in the interaction map to make the assignment of function. Hence, when we include alternate sources of information like BLAST interaction map, more information in addition to those inferred from Pubmed allows the prediction of more

function name which contributed to a rise in recall. This limitation of information from one source limits our improvement to the rise in precision in this project, and further highlight the contradiction between precision (less prediction) and recall (more prediction).

5. CONCLUSION

Organism species name and common name are useful in function prediction in Pubmed abstract analysis; this observation is useful in our case but is not used extensively in other work. Hence, maybe further work could be done to investigate this observation on other general text classification problem which was not done here. Furthermore, sentence contains more precise information which is useful in function prediction and is observed to be in use in other previous work. Lastly, the classification system uses the scoring function of protein pairs in **Equation 1** mainly for distributing the edges into different basket for prediction in those baskets, but without consideration to domain knowledge of information sources, such as high score in Pubmed co-occurrence have higher chance of function sharing while high BLAST e-value score would mean the opposite. Hence, further work could be done to incorporate domain knowledge into the prediction system to allow penalizing the function assignment score (**Equation 3**) for basket with low score in Pubmed for example or penalize basket with high BLAST e-value.

References:

(1) Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol*, 215: 403-410.

(2) Arakaki, A.K.; Zhang, Yang; Skolnick, Jeffrey (2004). Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics*, Vol. 20 no. 7, pages 1087–1096.

(3) Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25-9.

(4) Chiang, J.; Yu, H (2004). Extracting Functional Annotations of Proteins Based on Hybrid Text Mining Approaches. *Proc BioCreAtIvE Challenge Evaluation Workshop*.

(5) Chua Hon Nian, Sung Wing-Kin, Wong Limsoon (2007). An efficient strategy for extensive integration of diverse biological data for protein function prediction. *Bioinformatics*, 23: 3364 – 3373.

(6) Couto, F.M., Silva, M.J., and Coutinho, P.M. (2005). Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics*, 6(Suppl 1): S21.

(7) Fawcett, Tom (2004). ROC graphs: notes and practical considerations for researchers. *Kluwer Academic Publishers*, 12:56.

(8) Hirschman, L., Yeh, A., Blaschke, C. and Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1):S1.

(9) Letovsky, Stanley and Kasif, Simon (2003). Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, Vol. 19 Suppl. 1.

(10) Mitsumori T, Fation S, Murata M, Doi K, Doi H. (2005). Gene/protein name recognition based on support vector machine using dictionary as features. *BMC Bioinformatics*, 6 Suppl 1:S8. Epub.

(11) Ray, S. and Craven M. (2005). Learning statistical models for annotating proteins with function information using biomedical text. *BMC Bioinformatics*, 6(Suppl I):S18.

(12) Sjölander, Kimmen (2004). Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, *V*ol. 20 no. 2 2004, pages 170-179.

(13) Tamames J. (2005). Text detective: a rule-based system for gene annotation in biomedical texts. *BMC Bioinformatics*, 6 Suppl 1:S10. Epub.

(14) Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, Gilbert (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS*, vol. 99, no. 10, 6567-6572.

(15) Tusher, Tibshirani and Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98: 5116-5121.

(16) Verspoor, Karin; Cohn, Judith; Joslyn, Cliff; Mniszewski, Sue; Rechtsteiner, Andreas; Rocha, Luis M and Tiago Simas (2005). Protein annotation as term categorization in the gene ontology using word proximity networks. *BMC Bioinformatics*, 6(Suppl 1): S20.

APPENDIX A

Total abstract	Positive abstract	Negative abstract
192263	15600	176663

			Observed value				Expected value					
word	Word Index	Total	Positive (A)	Negative (B)	Not_word positive (C)	Not_word negative (D)	(A)	(B)	(C)	(D)	chi	OR
cerevisia	148	18637	4836	13801	10764	162862	1512.1849	17124.82	14087.82	159538.2	8804.403	0.724422
saccharomyc	157	18055	4684	13371	10916	163292	1464.9621	16590.04	14135.04	160073	8495.786	0.719354
yeast	27	27221	5908	21313	9692	155350	2208.6808	25012.32	13391.32	151650.7	7855.286	0.647694
defect	108	13697	2621	11076	12979	165587	1111.3589	12585.64	14488.64	164077.4	2402.925	0.479869
mutant	431	31411	5159	26252	10441	150411	2548.6526	28862.35	13051.35	147800.7	3477.806	0.451941
requir	411	30590	4662	25928	10938	150735	2482.0376	28107.96	13117.96	148555	2477.981	0.394079
homolog	151	12089	2003	10086	13597	166577	980.88764	11108.11	14619.11	165554.9	1236.892	0.386134
subunit	342	13899	2244	11655	13356	165008	1127.749	12771.25	14472.25	163891.7	1296.134	0.376339
complex	158	30757	4564	26193	11036	150470	2495.5878	28261.41	13104.41	148401.6	2221.051	0.375799
deletion	206	11830	1914	9916	13686	166747	959.87267	10870.13	14640.13	165792.9	1099.838	0.371387
strain	149	17407	2635	14772	12965	161891	1412.3841	15994.62	14187.62	160668.4	1266.463	0.347792
encod	155	27176	3795	23381	11805	153282	2205.0296	24970.97	13394.97	151692	1453.104	0.323774
mutat	594	28889	3981	24908	11619	151755	2344.0204	26544.98	13255.98	150118	1464.158	0.319628
essenti	85	17091	2442	14649	13158	162014	1386.7442	15704.26	14213.26	160958.7	959.1804	0.3123
protein	52	91523	9918	81605	5682	95058	7426.0716	84096.93	8173.928	92566.07	1736.824	0.308195
compon	192	14806	2118	12688	13482	163975	1201.3419	13604.66	14398.66	163058.3	824.7092	0.307556
allel	638	10401	1480	8921	14120	167742	843.92525	9557.075	14756.07	167105.9	551.5898	0.294655
suppress	126	11413	1609	9804	13991	166859	926.03777	10486.96	14673.96	166176	582.763	0.291654
phenotyp	71	16680	2257	14423	13343	162240	1353.3961	15326.6	14246.6	161336.4	718.9432	0.27938
interact	532	30141	3829	26312	11771	150351	2445.6063	27695.39	13154.39	148967.6	1009.971	0.269224
cycl	552	13795	1841	11954	13759	164709	1119.3105	12675.69	14480.69	163987.3	545.5513	0.265671
pathwai	383	25745	3298	22447	12302	154216	2088.9199	23656.08	13511.08	153006.9	879.3728	0.265246
function	122	54781	6377	48404	9223	128259	4444.8677	50336.13	11155.13	126326.9	1278.247	0.262951
conserv	506	14492	1905	12587	13695	164076	1175.8643	13316.14	14424.14	163346.9	532.1627	0.258456
gene	4	73609	8045	65564	7555	111099	5972.5501	67636.45	9627.45	109026.6	1268.153	0.256336
wild	133	16377	2121	14256	13479	162407	1328.8111	15048.19	14271.19	161614.8	561.8351	0.25349
recombin	1590	13925	1816	12109	13784	164554	1129.8586	12795.14	14470.14	163867.9	488.8832	0.252941
genet	142	17848	2218	15630	13382	161033	1448.1663	16399.83	14151.83	160263.2	490.9501	0.232396
kinas	219	22812	2780	20032	12820	156631	1850.9396	20961.06	13749.06	155701.9	575.8341	0.22931
lack	484	12713	1594	11119	14006	165544	1031.5183	11681.48	14568.48	164981.5	357.4377	0.229022
regulatori	724	10309	1278	9031	14322	167632	836.46047	9472.54	14763.54	167190.5	268.0266	0.219148
propos	736	11198	1383	9815	14217	166848	908.59292	10289.41	14691.41	166373.6	286.2492	0.218445
chromosom	75	15279	1853	13426	13747	163237	1239.7206	14039.28	14360.28	162623.7	358.678	0.214539
dna	234	34403	3963	30440	11637	146223	2791.4201	31611.58	12808.58	145051.4	651.767	0.213754
overexpress	287	11547	1388	10159	14212	166504	936.91038	10610.09	14663.09	166052.9	251.4646	0.204308
absenc	842	10157	1224	8933	14376	167730	824.12737	9332.873	14775.87	167330.1	222.9311	0.203757
act	515	12105	1449	10656	14151	166007	982.18586	11122.81	14617.81	165540.2	257.6835	0.202813
sensit	235	19820	2320	17500	13280	159163	1608.1721	18211.83	13991.83	158451.2	382.3118	0.201094
mammalian	2110	10079	1204	8875	14396	167788	817.79854	9261.201	14782.2	167401.8	209.4677	0.198977
genom	15	14284	1678	12606	13922	164057	1158.9874	13125.01	14441.01	163538	273.2462	0.195508
growth	87	25341	2885	22456	12715	154207	2056.1398	23284.86	13543.86	153378.1	418.8343	0.192601
domain	37	23651	2637	21014	12963	155649	1919.0151	21731.98	13680.98	154931	333.357	0.178042
phosphoryl	545	13714	1553	12161	14047	164502	1112.7383	12601.26	14487.26	164061.7	204.1349	0.17479
transcript	33	31932	3475	28457	12125	148206	2590.926	29341.07	13009.07	147321.9	393.6866	0.17395
regul	230	44522	4721	39801	10879	136862	3612.4642	40909.54	11987.54	135753.5	481.7711	0.173834
rna	145	14878	1669	13209	13931	163454	1207.1839	13670.82	14392.82	162992.2	208.398	0.171
distinct	712	12346	1386	10960	14214	165703	1001.7403	11344.26	14598.26	165318.7	171.4226	0.168567
promot	559	22183	2437	19746	13163	156917	1799.9033	20383.1	13800.1	156279.9	277.4305	0.167692
doe	81	10475	1178	9297	14422	167366	849.92952	9625.07	14750.07	167037.9	145.7579	0.167444
independ	605	14622	1605	13017	13995	163646	1186.4124	13435.59	14413.59	163227.4	173.9561	0.158897
depend	45	40890	4246	36644	11354	140019	3317.7678	37572.23	12282.23	139090.8	358.9754	0.155015
identifi	60	37559	3914	33645	11686	143018	3047.4943	34511.51	12552.51	142151.5	333.2301	0.153425
role	410	43495	4448	39047	11152	137616	3529.1346	39965.87	12070.87	136697.1	336.4898	0.147893
involv	201	36459	3768	32691	11832	143972	2958.2416	33500.76	12641.76	143162.2	297.6765	0.146903
synthesi	571	11602	1235	10367	14365	166296	941.37302	10660.63	14658.63	166002.4	106.0746	0.13959
contain	18	36438	3720	32718	11880	143945	2956.5377	33481.46	12643.46	143181.5	264.7284	0.139136
previous	322	20185	2116	18069	13484	158594	1637.7878	18547.21	13962.21	158115.8	169.7869	0.13905
togeth	428	11482	1221	10261	14379	166402	931.63635	10550.36	14668.36	166112.6	104.0242	0.138956
termin	120	21536	2252	19284	13348	157379	1747.4064	19788.59	13852.59	156874.4	178.5801	0.138902
amino	156	22001	2268	19733	13332	156930	1785.136	20215.86	13814.86	156447.1	160.5116	0.131261
substrat	1919	11993	1251	10742	14349	165921	973.09831	11019.9	14626.9	165643.1	92.11875	0.129252
phase	815	11462	1196	10266	14404	166397	930.01358	10531.99	14669.99	166131	88.03891	0.128992
ident	315	10693	1115	9578	14485	167085	867.61779	9825.382	14732.38	166837.6	81.28496	0.128019

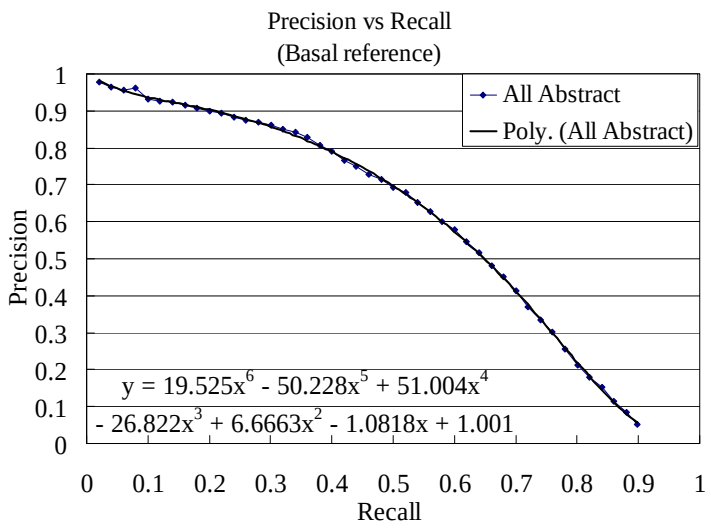
			Observed value				Expected value					
word	Word Index	Total	Positive (A)	Negative (B)	Not_word positive (C)	Not_word negative (D)	(A)	(B)	(C)	(D)	chi	OR
bind	1099	42643	4224	38419	11376	138244	3460.0043	39183	12140	137480	235.9182	0.125835
format	172	17277	1760	15517	13840	161146	1401.836	15875.16	14198.16	160787.8	109.4231	0.120788
structur	453	28173	2820	25353	12780	151310	2285.925	25887.07	13314.07	150775.9	159.1132	0.119557
nucleotid	1122	11346	1160	10186	14440	166477	920.60147	10425.4	14679.4	166237.6	72.00087	0.118241
vivo	491	18578	1876	16702	13724	159961	1507.3977	17070.6	14092.6	159592.4	108.5855	0.116998
suggest	207	66022	6235	59787	9365	116876	5356.9496	60665.05	10243.05	115997.9	238.543	0.114447
accumul	627	12282	1245	11037	14355	165626	996.54744	11285.45	14603.45	165377.5	72.01254	0.114444
form	400	28068	2783	25285	12817	151378	2277.4054	25790.59	13322.59	150872.4	143.0376	0.113927
activ	107	80992	7462	73530	8138	103133	6571.5983	74420.4	9028.402	102242.6	226.8636	0.109271
process	831	22298	2208	20090	13392	156573	1809.2342	20488.77	13790.77	156174.2	108.2	0.108891
effici	124	10746	1071	9675	14529	166988	871.91815	9874.082	14728.08	166788.9	52.39817	0.104588
transport	1223	11688	1159	10529	14441	166134	948.35096	10739.65	14651.65	165923.4	54.21733	0.102558
initi	817	15175	1498	13677	14102	162986	1231.2821	13943.72	14368.72	162719.3	68.26581	0.102391
loss	454	11245	1113	10132	14487	166531	912.40644	10332.59	14687.59	166330.4	50.97646	0.101317
implic	1111	10165	1007	9158	14593	167505	824.77648	9340.224	14775.22	167322.8	46.26081	0.101112
plai	584	19867	1942	17925	13658	158738	1611.9857	18255.01	13988.01	158408	82.00174	0.100084
show	12	65409	6067	59342	9533	117321	5307.2115	60101.79	10292.79	116561.2	179.4158	0.099758
known	144	20420	1987	18433	13613	158230	1656.8555	18763.14	13943.14	157899.9	80.10094	0.097937
lead	647	14642	1423	13219	14177	163444	1188.0351	13453.96	14411.96	163209	54.74293	0.093791
enzym	6	21672	2084	19588	13516	157075	1758.4413	19913.56	13841.56	156749.4	73.92996	0.092167
residu	855	15969	1542	14427	14058	162236	1295.7064	14673.29	14304.29	161989.7	55.56585	0.091132
addition	339	32897	3119	29778	12481	146885	2669.225	30227.78	12930.78	146435.2	99.50747	0.090848
nuclear	982	14595	1408	13187	14192	163476	1184.2216	13410.78	14415.78	163252.2	49.80121	0.089867
consist	267	18285	1751	16534	13849	160129	1483.624	16801.38	14116.38	159861.6	57.95254	0.08796
condition	248	15542	1483	14059	14117	162604	1261.0601	14280.94	14338.94	162382.1	46.24796	0.084576
singl	173	18583	1764	16819	13836	159844	1507.8034	17075.2	14092.2	159587.8	52.44428	0.083384
mediat	501	30396	2849	27547	12751	149116	2466.2967	27929.7	13133.7	148733.3	76.7656	0.082598
sequenc	24	37169	3461	33708	12139	142955	3015.8502	34153.15	12584.15	142509.9	88.64485	0.082485
exhibit	42	14517	1380	13137	14220	163526	1177.8928	13339.11	14422.11	163323.9	40.8229	0.08207
evid	629	16320	1548	14772	14052	161891	1324.1861	14995.81	14275.81	161667.2	44.98821	0.081816
factor	439	37423	3469	33954	12131	142709	3036.4594	34386.54	12563.54	142276.5	83.26239	0.079868
target	253	20973	1972	19001	13628	157662	1701.7252	19271.27	13898.27	157391.7	52.4367	0.079425
thu	352	21200	1990	19210	13610	157453	1720.1438	19479.86	13879.86	157183.1	51.78332	0.078619
membran	183	19605	1839	17766	13761	158897	1590.7273	18014.27	14009.27	158648.7	46.95927	0.077458
affect	112	19385	1818	17567	13782	159096	1572.8767	17812.12	14027.12	158850.9	46.23602	0.077243
local	198	18566	1741	16825	13859	159838	1506.424	17059.58	14093.58	159603.4	44.0021	0.076792
similar	332	28499	2644	25855	12956	150808	2312.3763	26186.62	13287.62	150476.4	60.76592	0.07567
hybrid	276	11068	1041	10027	14559	166636	898.04487	10169.96	14701.96	166493	26.27853	0.074917
member	713	11768	1104	10664	14496	165999	954.84207	10813.16	14645.16	165849.8	27.01107	0.073906
import	445	27499	2525	24974	13075	151689	2231.2374	25267.76	13368.76	151395.2	49.11687	0.069286
furthermor	425	14104	1308	12796	14292	163867	1144.3824	12959.62	14455.62	163703.4	27.47431	0.068932
caus	210	22912	2105	20807	13495	155856	1859.0535	21052.95	13740.95	155610.1	40.20199	0.067593
thei	617	14909	1374	13535	14226	163128	1209.6992	13699.3	14390.3	162963.7	26.32732	0.065974
isol	351	23042	2106	20936	13494	155727	1869.6015	21172.4	13730.4	155490.6	36.95999	0.064788
support	911	12043	1108	10935	14492	165728	977.15525	11065.84	14622.84	165597.2	20.34192	0.063989
abil	570	12107	1111	10996	14489	165667	982.34814	11124.65	14617.65	165538.3	19.56878	0.062677
map	502	10896	995	9901	14605	166762	884.08898	10011.91	14715.91	166651.1	16.05244	0.05974
site	522	28909	2605	26304	12995	150359	2345.6432	26563.36	13254.36	150099.6	36.7324	0.059139
cell	166	93919	8107	85812	7493	90851	7620.4803	86298.52	7979.52	90364.48	66.08707	0.058986
cellular	1884	15638	1419	14219	14181	162444	1268.8494	14369.15	14331.15	162293.8	21.04929	0.05811
indic	295	43791	3900	39891	11700	136772	3553.1517	40237.85	12046.85	136425.2	47.7163	0.058001
elem	377	10039	914	9125	14686	167538	814.55298	9224.447	14785.45	167438.6	13.94134	0.057923
multipl	834	11194	1016	10178	14584	166485	908.26836	10285.73	14691.73	166377.3	14.76638	0.05673
product	897	23246	2093	21153	13507	155510	1886.1539	21359.85	13713.85	155303.2	28.08233	0.056597
sever	347	26720	2399	24321	13201	152342	2168.0303	24551.97	13431.97	152111	31.1014	0.056262
novel	622	16921	1527	15394	14073	161269	1372.9506	15548.05	14227.05	161115	20.62648	0.055652
link	494	13120	1186	11934	14414	164729	1064.5418	12055.46	14535.46	164607.5	16.1859	0.055284
part	821	10029	906	9123	14694	167540	813.74159	9215.258	14786.26	167447.7	12.00997	0.053969
type	152	40049	3547	36502	12053	140161	3249.5301	36799.47	12350.47	139863.5	37.43319	0.053077
specif	38	48508	4268	44240	11332	132423	3935.8837	44572.12	11664.12	132090.9	40.79069	0.052066
dure	406	29018	2576	26442	13024	150221	2354.4873	26663.51	13245.51	149999.5	26.71202	0.050638
shown	544	21558	1914	19644	13686	157019	1749.1915	19808.81	13850.81	156854.2	19.03363	0.048388
mani	525	11088	987	10101	14613	166562	899.66764	10188.33	14700.33	166474.7	9.790747	0.046789
signal	521	26238	2316	23922	13284	152741	2128.9213	24109.08	13471.08	152553.9	20.71864	0.046568
highli	561	14542	1289	13253	14311	163410	1179.9213	13362.08	14420.08	163300.9	11.87229	0.045547
clone	8	15714	1391	14323	14209	162340	1275.016	14438.98	14324.98	162224	12.50435	0.045155
wherea	118	20977	1851	19126	13749	157537	1702.0498	19274.95	13897.95	157388	15.92333	0.044894
gener	833	22430	1974	20456	13626	156207	1819.9446	20610.06	13780.06	156052.9	16.06644	0.043859

			Observed value				Expected value					
word	Word Index	Total	Positive (A)	Negative (B)	Not_word positive (C)	Not_word negative (D)	(A)	(B)	(C)	(D)	chi	OR
appear	455	17043	1500	15543	14100	161120	1382.8495	15660.15	14217.15	161002.8	11.85155	0.042487
find	588	19997	1750	18247	13850	158416	1622.5337	18374.47	13977.47	158288.5	12.16307	0.040196
contrast	601	17561	1531	16030	14069	160633	1424.8795	16136.12	14175.12	160526.9	9.466048	0.037613
found	34	41933	3616	38317	11984	138346	3402.3957	38530.6	12197.6	138132.4	18.66532	0.037202
vitro	39	22379	1937	20442	13663	156221	1815.8065	20563.19	13784.19	156099.8	9.962828	0.0348
fold	562	13631	1180	12451	14420	164212	1106.0038	12525	14494	164138	5.798949	0.033117
direct	683	15556	1344	14212	14256	162451	1262.1961	14293.8	14337.8	162369.2	6.277891	0.032469
famili	1444	21069	1815	19254	13785	157409	1709.5146	19359.49	13890.49	157303.5	7.955536	0.031978
reveal	396	27894	2394	25500	13206	151163	2263.2873	25630.71	13336.71	151032.3	9.60997	0.031258
mrna	388	17907	1540	16367	14060	160296	1452.9535	16454.05	14147.05	160209	6.258349	0.030489
same	32	14729	1268	13461	14332	163202	1195.0942	13533.91	14404.91	163129.1	5.241868	0.030461
acid	159	33882	2893	30989	12707	145674	2749.1467	31132.85	12850.85	145530.1	9.944529	0.02948
character	338	22504	1921	20583	13679	156080	1825.9488	20678.05	13774.05	155984.9	6.098729	0.027312
includ	88	29272	2488	26784	13112	149879	2375.0966	26896.9	13224.9	149766.1	6.889931	0.026047
associ	272	40894	3464	37430	12136	139233	3318.0924	37575.91	12281.91	139087.1	8.869032	0.026025
report	429	31497	2661	28836	12939	147827	2555.6306	28941.37	13044.37	147721.6	5.654352	0.022963
thi	14	99983	8277	91706	7323	84957	8112.5063	91870.49	7487.494	84792.51	7.562784	0.019985
region	256	30469	2558	27911	13042	148752	2472.2198	27996.78	13127.78	148666.2	3.849198	0.019244
level	262	42620	3322	39298	12278	137365	3458.1381	39161.86	12141.86	137501.1	7.493872	-0.02422
molecular	804	20960	1624	19336	13976	157327	1700.6704	19259.33	13899.33	157403.7	4.221984	-0.02436
normal	604	20606	1589	19017	14011	157646	1671.9473	18934.05	13928.05	157728.9	5.016098	-0.0268
major	1082	17996	1379	16617	14221	160046	1460.1749	16535.83	14139.83	160127.2	5.418374	-0.02967
new	141	16048	1228	14820	14372	161843	1302.1164	14745.88	14297.88	161917.1	5.009351	-0.03007
differ	56	43871	3378	40493	12222	136170	3559.6428	40311.36	12040.36	136351.6	13.06969	-0.03178
model	1611	19830	1508	18322	14092	158341	1608.9835	18221.02	13991.02	158442	7.69086	-0.03395
induct	414	10715	812	9903	14788	166760	869.40285	9845.597	14730.6	166817.4	4.368177	-0.03403
result	80	84119	6547	77572	9053	99091	6825.3195	77293.68	8774.681	99369.32	21.95875	-0.03442
earli	477	11606	879	10727	14721	165936	941.69757	10664.3	14658.3	165998.7	4.834828	-0.03449
respect	466	26032	1982	24050	13618	152613	2112.2067	23919.79	13487.79	152743.2	10.10332	-0.03453
non	556	14753	1111	13642	14489	163021	1197.0416	13555.96	14402.96	163107	7.290046	-0.03796
further	312	17854	1346	16508	14254	160155	1448.6531	16405.35	14151.35	160257.7	8.726841	-0.03805
variou	260	12509	937	11572	14663	165091	1014.966	11494.03	14585.03	165169	6.971494	-0.04017
high	237	30941	2329	28612	13271	148051	2510.5174	28430.48	13089.48	148232.5	17.02258	-0.04187
observ	170	34791	2620	32171	12980	144492	2822.902	31968.1	12777.1	144694.9	19.37846	-0.04259
chain	793	13204	984	12220	14616	164443	1071.3575	12132.64	14528.64	164530.4	8.323679	-0.04289
recent	580	16490	1223	15267	14377	161396	1337.9797	15152.02	14262.02	161511	11.76215	-0.0461
possibl	941	16090	1189	14901	14411	161762	1305.5242	14784.48	14294.48	161878.5	12.35247	-0.04785
assai	117	18565	1373	17192	14227	159471	1506.3429	17058.66	14093.66	159604.3	14.21893	-0.04809
present	58	32917	2440	30477	13160	146186	2670.8477	30246.15	12929.15	146416.8	26.20033	-0.05093
express	7	69104	5204	63900	10396	112763	5607.0196	63496.98	9992.98	113166	49.21527	-0.05386
due	609	11806	852	10954	14748	165709	957.92534	10848.07	14642.07	165814.9	13.58126	-0.05852
repres	254	10247	732	9515	14868	167148	831.42986	9415.57	14768.57	167247.4	13.66924	-0.06305
approxim	661	11631	831	10800	14769	165863	943.72604	10687.27	14656.27	165975.7	15.59745	-0.06342
analyz	286	12422	885	11537	14715	165126	1007.9069	11414.09	14592.09	165248.9	17.4377	-0.06509
confirm	344	12582	894	11688	14706	164975	1020.8891	11561.11	14579.11	165101.9	18.36596	-0.06648
mai	528	44057	3196	40861	12404	135802	3574.7346	40482.27	12025.27	136180.7	56.65082	-0.06736
rate	1264	15017	1060	13957	14540	162706	1218.4622	13798.54	14381.54	162864.5	24.32813	-0.07065
rel	285	11496	809	10687	14791	165976	932.7723	10563.23	14667.23	166099.8	19.01069	-0.07086
base	637	21627	1521	20106	14079	156557	1754.7901	19872.21	13845.21	156790.8	38.19461	-0.0751
pattern	171	12390	862	11528	14738	165135	1005.3104	11384.69	14594.69	165278.3	23.76486	-0.07685
number	1160	16469	1136	15333	14464	161330	1336.2758	15132.72	14263.72	161530.3	35.72751	-0.08282
system	1240	24382	1686	22696	13914	153967	1978.3276	22403.67	13621.67	154259.3	53.8376	-0.08512
molecul	1321	12337	843	11494	14757	165169	1001.0101	11335.99	14598.99	165327	29.00568	-0.08571
induc	212	44381	3088	41293	12512	135370	3601.0236	40779.98	11998.98	135883	103.414	-0.092
reaction	1515	13044	873	12171	14727	164492	1058.3752	11985.62	14541.62	164677.4	37.90754	-0.09628
enhanc	533	17086	1140	15946	14460	160717	1386.3385	15699.66	14213.66	160963.3	52.28343	-0.09985
low	797	17885	1191	16694	14409	159969	1451.1685	16433.83	14148.83	160229.2	55.96875	-0.10125
select	530	18339	1218	17121	14382	159542	1488.0055	16850.99	14111.99	159812	58.94229	-0.10283
design	641	10797	711	10086	14889	166577	876.05624	9920.944	14723.94	166742.1	35.8577	-0.1031
chang	541	21669	1436	20233	14164	156430	1758.1979	19910.8	13841.8	156752.2	72.4202	-0.10577
increas	203	48623	3308	45315	12292	131348	3945.2146	44677.79	11654.79	131985.2	149.924	-0.10788
speci	687	11118	721	10397	14879	166266	902.10181	10215.9	14697.9	166447.1	41.99615	-0.11074
detect	93	24954	1631	23323	13969	153340	2024.739	22929.26	13575.26	153733.7	95.75785	-0.11484
beta	672	19326	1241	18085	14359	158578	1568.0895	17757.91	14031.91	158905.1	82.55063	-0.12043
higher	304	17783	1139	16644	14461	160019	1442.8923	16340.11	14157.11	160322.9	76.75481	-0.12076
inhibit	578	32600	2122	30478	13478	146185	2645.1267	29954.87	12954.87	146708.1	135.5841	-0.12196
determin	516	31935	2075	29860	13525	146803	2591.1694	29343.83	13008.83	147319.2	134.1915	-0.12247
examin	448	23085	1474	21611	14126	155052	1873.0905	21211.91	13726.91	155451.1	105.1686	-0.12572

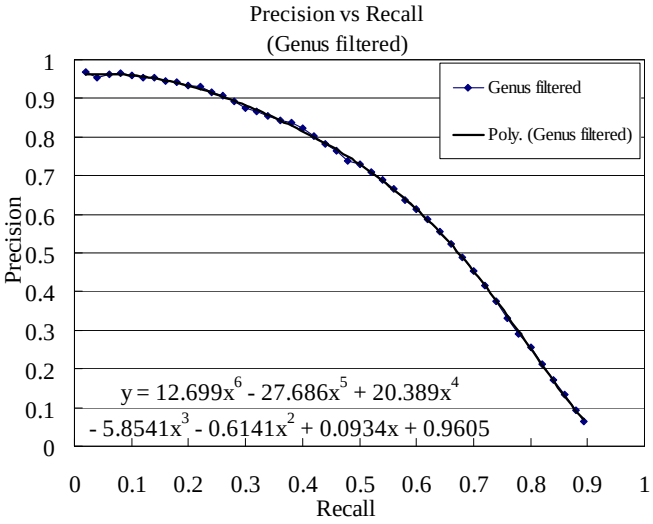
			Observed value				Expected value					
word	Word Index	Total	Positive (A)	Negative (B)	Not_word positive (C)	Not_word negative (D)	(A)	(B)	(C)	(D)	chi	OR
human	136	40038	2607	37431	12993	139232	3248.6375	36789.36	12351.36	139873.6	174.196	-0.12706
group	36	20694	1309	19385	14291	157278	1679.0875	19014.91	13920.91	157648.1	99.48154	-0.12892
less	1038	10932	682	10250	14918	166413	887.00998	10044.99	14712.99	166618	54.67582	-0.12946
resist	223	13853	867	12986	14733	163677	1124.0166	12728.98	14475.98	163934	68.92491	-0.12976
produc	474	16519	1037	15482	14563	161181	1340.3328	15178.67	14259.67	161484.3	81.73186	-0.12999
stimul	846	18818	1184	17634	14416	159029	1526.871	17291.13	14073.13	159371.9	92.88449	-0.13037
surfac	354	10172	622	9550	14978	167113	825.34445	9346.656	14774.66	167316.3	57.56874	-0.13866
obtain	681	14159	863	13296	14737	163367	1148.8451	13010.15	14451.15	163652.8	83.55494	-0.14295
correl	271	14141	856	13285	14744	163378	1147.3846	12993.62	14452.62	163669.4	86.92653	-0.14631
decreas	240	23273	1424	21849	14176	154814	1888.3446	21384.66	13711.66	155278.3	141.3788	-0.14767
differenti	379	15561	941	14620	14659	162043	1262.6017	14298.4	14337.4	162364.6	97.00068	-0.14783
follow	640	19599	1172	18427	14428	158236	1590.2405	18008.76	14009.76	158654.2	133.301	-0.15643
revers	1318	10515	616	9899	14984	166764	853.17508	9661.825	14746.82	167001.2	75.906	-0.15954
potenti	723	19517	1157	18360	14443	158303	1583.5871	17933.41	14016.41	158729.6	139.191	-0.16071
posit	162	13801	809	12992	14791	163671	1119.7974	12681.2	14480.2	163981.8	101.1382	-0.16175
test	64	18987	1120	17867	14480	158796	1540.5835	17446.42	14059.42	159216.6	138.6521	-0.16276
case	268	12094	704	11390	14896	165273	981.29333	11112.71	14618.71	165550.3	91.00092	-0.16382
deriv	1516	15601	910	14691	14690	161972	1265.8473	14335.15	14334.15	162327.8	118.481	-0.16559
perform	772	13678	793	12885	14807	163778	1109.8173	12568.18	14490.18	164094.8	105.9661	-0.16702
effect	979	47597	2896	44701	12704	131962	3861.9662	43735.03	11738.03	132928	349.4578	-0.17201
signific	980	24352	1421	22931	14179	153732	1975.8934	22376.11	13624.11	154286.9	194.188	-0.17271
total	389	10257	578	9679	15022	166984	832.24125	9424.759	14767.76	167238.2	89.29002	-0.17796
investig	218	28660	1667	26993	13933	149670	2325.4396	26334.56	13274.56	150328.4	238.4413	-0.17823
concentr	565	18088	1021	17067	14579	159596	1467.6396	16620.36	14132.36	160042.6	163.2883	-0.18384
rang	990	10085	554	9531	15046	167132	818.28537	9266.715	14781.71	167396.3	98.0373	-0.18999
lower	1238	11931	656	11275	14944	165388	968.0677	10962.93	14631.93	165700.1	116.7253	-0.19118
cancer	366	14209	782	13427	14818	163236	1152.902	13056.1	14447.1	163606.9	140.2233	-0.19275
studi	427	71963	4393	67570	11207	109093	5838.9955	66124	9761.004	110539	622.8392	-0.19869
marker	86	10111	544	9567	15056	167096	820.39498	9290.605	14779.61	167372.4	106.9668	-0.19992
peptid	353	14355	770	13585	14830	163078	1164.7483	13190.25	14435.25	163472.7	157.3471	-0.20531
inhibitor	573	19038	1026	18012	14574	158651	1544.7216	17493.28	14055.28	159169.7	210.4038	-0.20755
time	370	20821	1114	19707	14486	156956	1689.3921	19131.61	13910.61	157531.4	239.1806	-0.21291
primari	1601	10438	546	9892	15054	166771	846.92739	9591.073	14753.07	167071.9	123.0466	-0.21362
compar	290	30172	1602	28570	13998	148093	2448.1216	27723.88	13151.88	148939.1	377.5022	-0.22678
pcr	232	10986	559	10427	15041	166236	891.39148	10094.61	14708.61	166568.4	143.0653	-0.2273
antibodi	848	12925	657	12268	14943	164395	1048.7197	11876.28	14551.28	164786.7	170.7124	-0.22976
prolifer	390	10240	510	9730	15090	166933	830.86189	9409.138	14769.14	167253.9	142.4384	-0.23669
develop	298	26978	1386	25592	14214	151071	2188.9641	24789.04	13411.04	151874	372.8773	-0.23988
measur	357	15668	751	14917	14849	161746	1271.2836	14396.72	14328.72	162266.3	252.293	-0.26091
mous	1610	12540	587	11953	15013	164710	1017.4813	11522.52	14582.52	165140.5	212.0432	-0.26859
cultur	1245	13415	623	12792	14977	163871	1088.4778	12326.52	14511.52	164336.5	232.8842	-0.27337
significantli	2147	23309	1080	22229	14520	154434	1891.2656	21417.73	13708.73	155245.3	430.9739	-0.28672
tissu	306	18405	837	17568	14763	159095	1493.3607	16911.64	14106.64	159751.4	347.1935	-0.28951
diseas	1605	16785	724	16061	14876	160602	1361.9157	15423.08	14238.08	161239.9	356.2866	-0.31277
treatment	1728	21582	913	20669	14687	155994	1751.1388	19830.86	13848.86	156832.1	491.7813	-0.32867
conclusion	534	20139	828	19311	14772	157352	1634.0554	18504.94	13965.94	158158.1	483.3564	-0.34034
method	671	23063	948	22115	14652	154548	1871.3055	21191.69	13728.69	155471.3	563.3672	-0.34471
treat	2206	11303	441	10862	15159	165801	917.1125	10385.89	14682.89	166277.1	285.7984	-0.35255
line	358	16797	665	16132	14935	160531	1362.8894	15434.11	14237.11	161228.9	426.1529	-0.35351
receptor	1135	28060	1139	26921	14461	149742	2276.7563	25783.24	13323.24	150879.8	724.5139	-0.35842
tumor	250	15232	572	14660	15028	162003	1235.9071	13996.09	14364.09	162666.9	421.5269	-0.37612
rat	2115	16742	591	16151	15009	160512	1358.4267	15383.57	14241.57	161279.4	516.8379	-0.40746
mice	1008	14271	494	13777	15106	162886	1157.9326	13113.07	14442.07	163549.9	447.5176	-0.41269
evalu	536	12427	420	12007	15180	164656	1008.3126	11418.69	14591.69	165244.3	399.3836	-0.42088
clinic	3170	11641	377	11264	15223	165399	944.53743	10696.46	14655.46	165966.5	395.0436	-0.43932
patient	1995	20486	640	19846	14960	156817	1662.2106	18823.79	13937.79	157839.2	765.7298	-0.47103
dai	1898	10337	278	10059	15322	166604	838.73236	9498.268	14761.27	167164.7	431.1604	-0.52214
dose	2399	10387	275	10112	15325	166551	842.7893	9544.211	14757.21	167118.8	440.0741	-0.52936

APPENDIX B

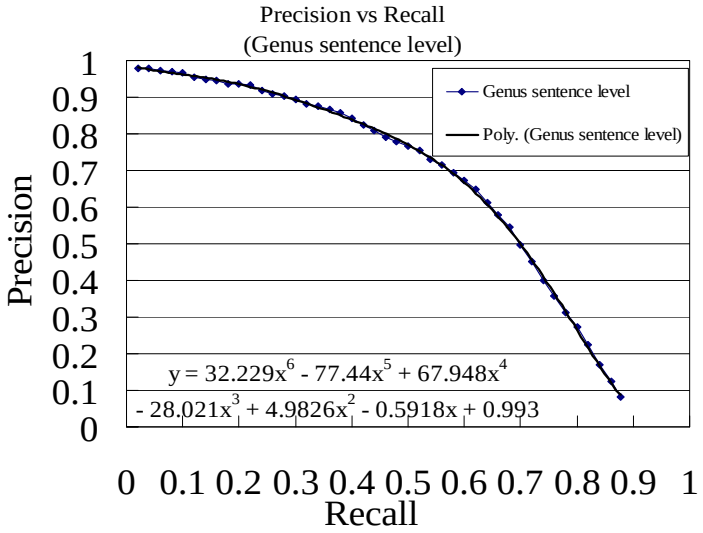
Note: All the graphs presented here are for reference for the calculation of the AUPRC
Hence, only the graph source is labelled
"Poly. (Graph)" is the line estimated by Microsoft Excel with a 6th polynomial equation of the original graph for calculation of AUC



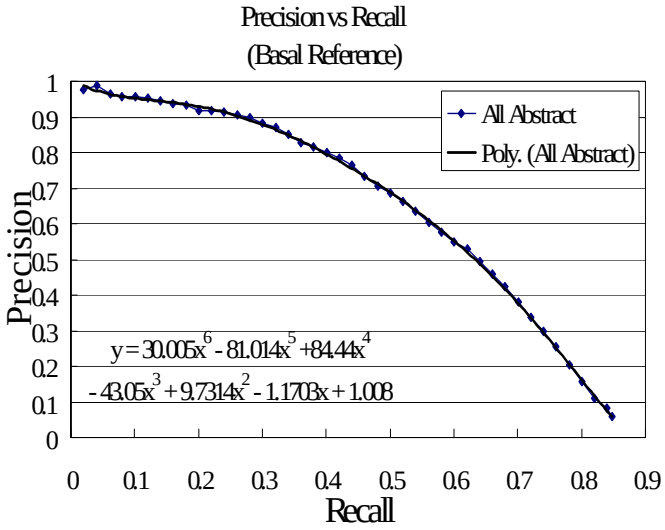
Part 1. Basal reference: Biological Process



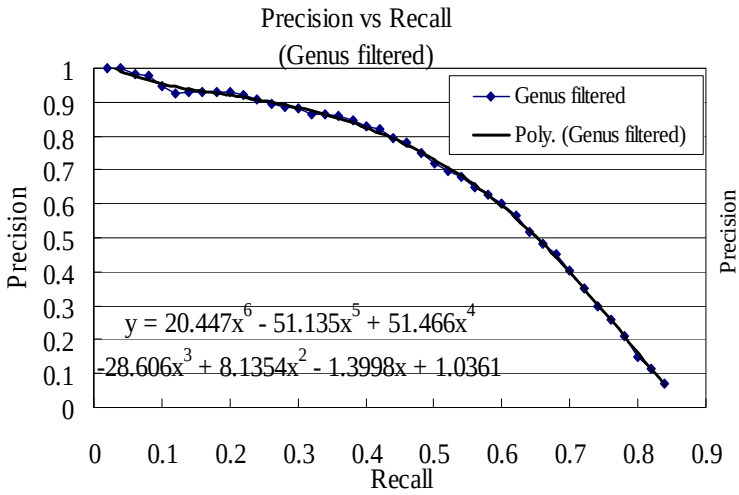
Part 2. Genus Filtered: Biological Process



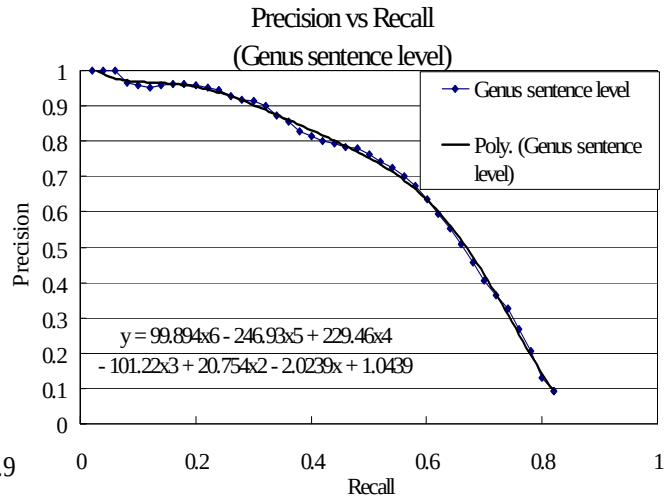
Part 3. Genus sentence level: Biological Process



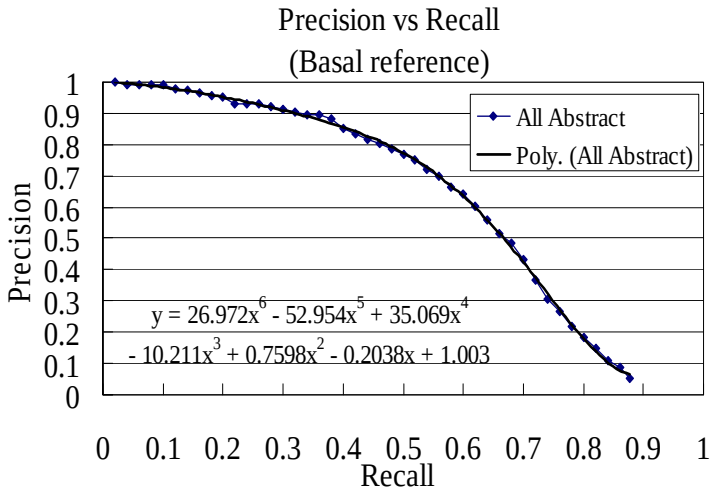
Part 4. Basal Reference: Cellular Component



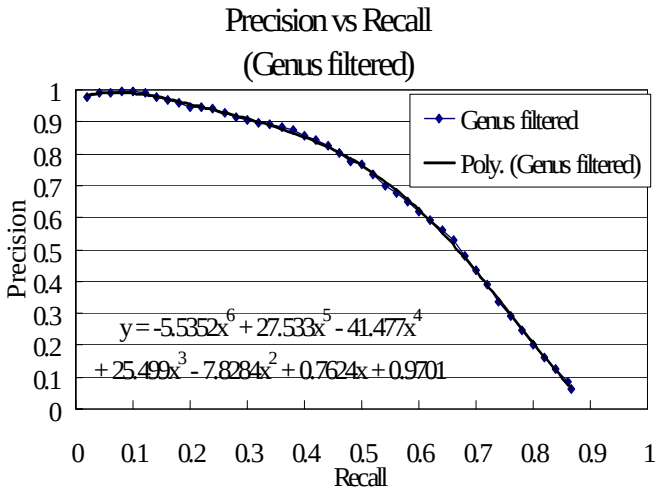
Part 5. Genus filtered: Cellular Component



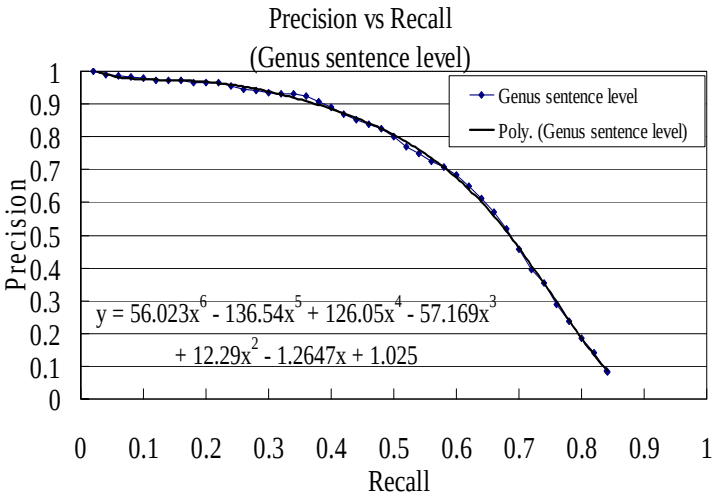
Part 6. Genus Sentence level: Cellular Component



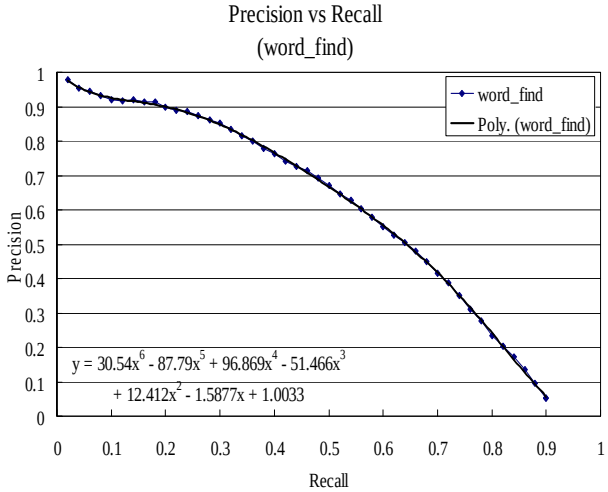
Part 7. Basal reference: Molecular Function



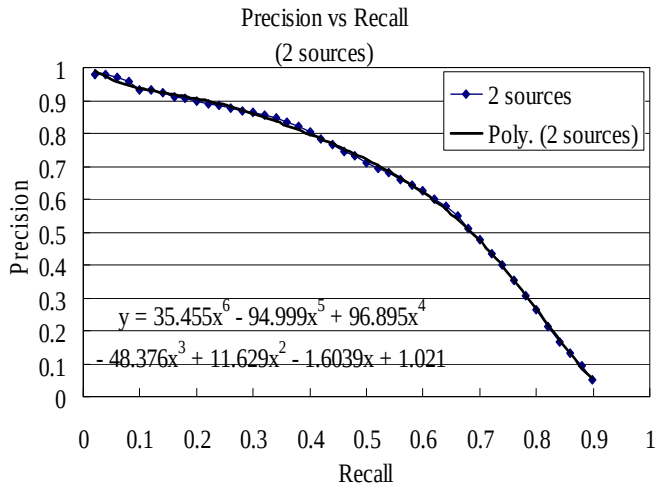
Part 8. Genus Filtered: Molecular Function



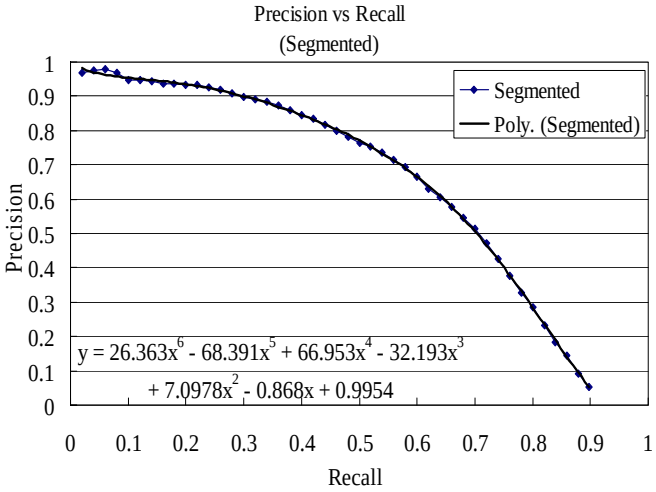
Part 9. Genus Sentence level: Molecular Function



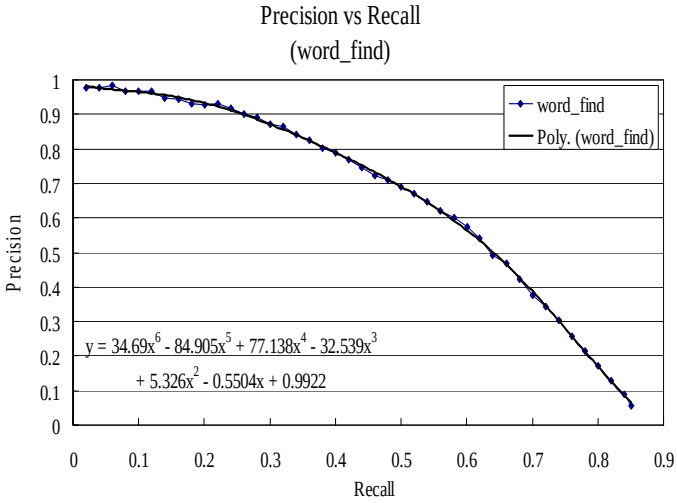
Part 10. Word_find: Biological Process



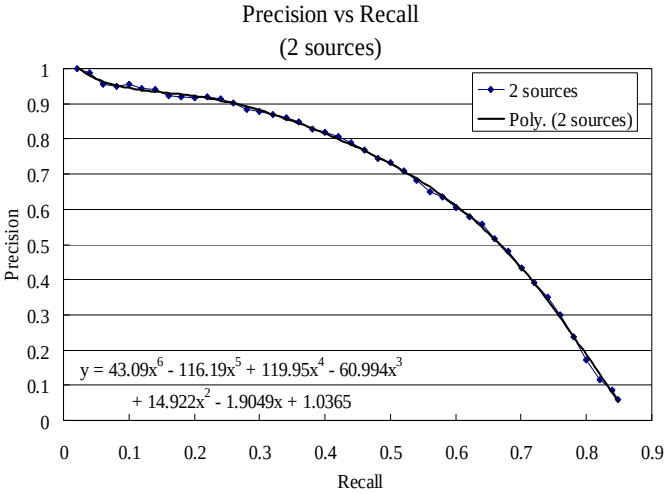
Part 11. 2 Sources: Biological Process



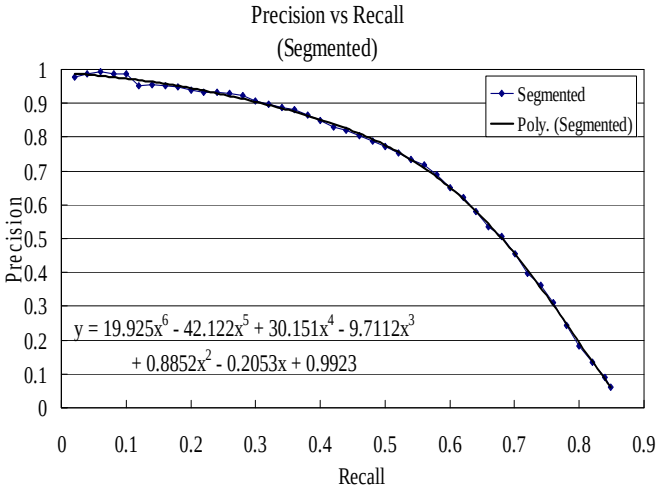
Part 12. Segmented: Biological Process



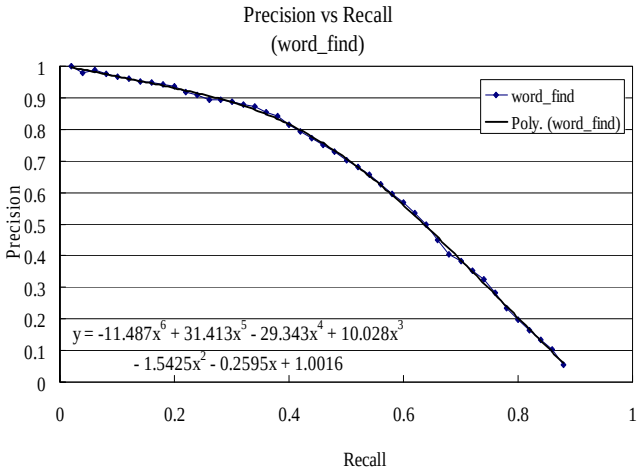
Part 13. Word_find: Cellular Component



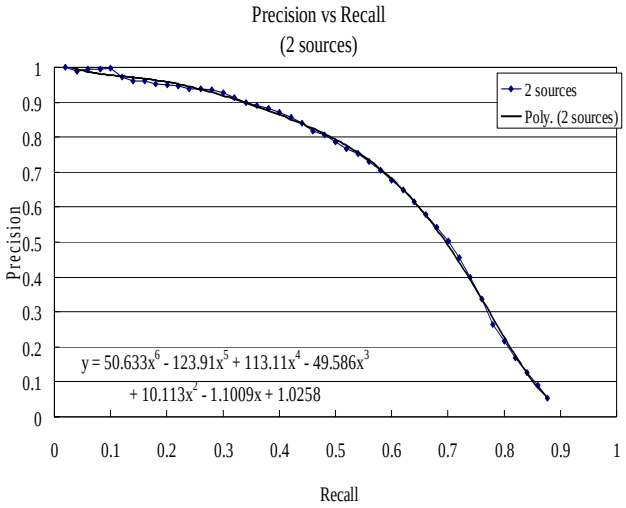
Part 14. 2 Sources: Cellular Component



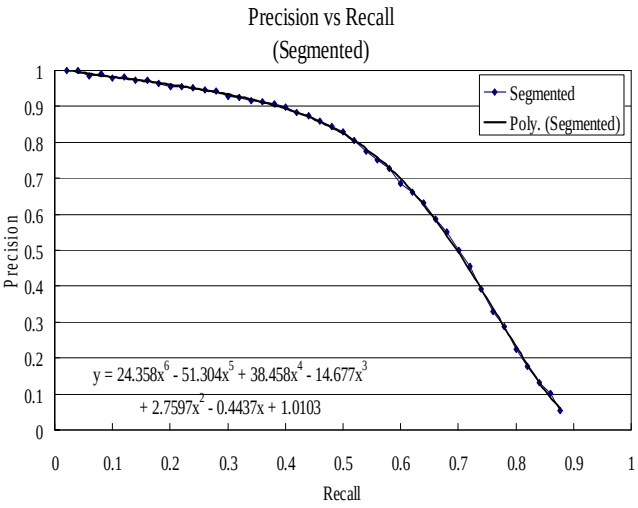
Part 15. Segmented: Cellular Component



Part 16. Word_find: Molecular Function

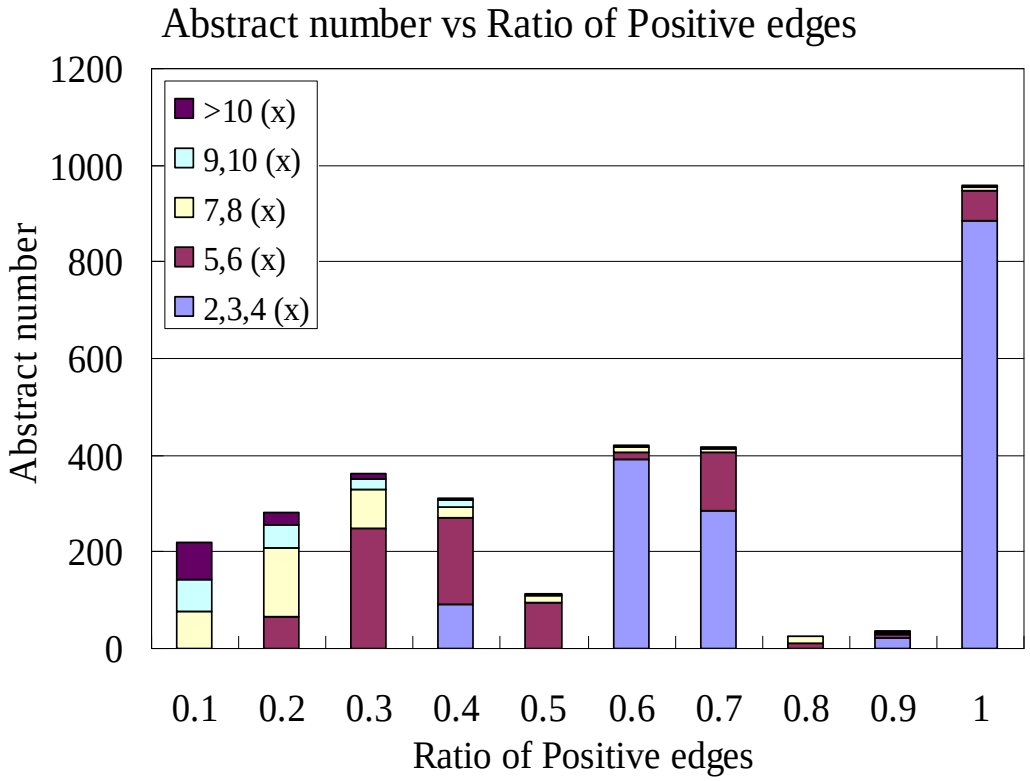


Part 17. 2 Sources: Molecular Function

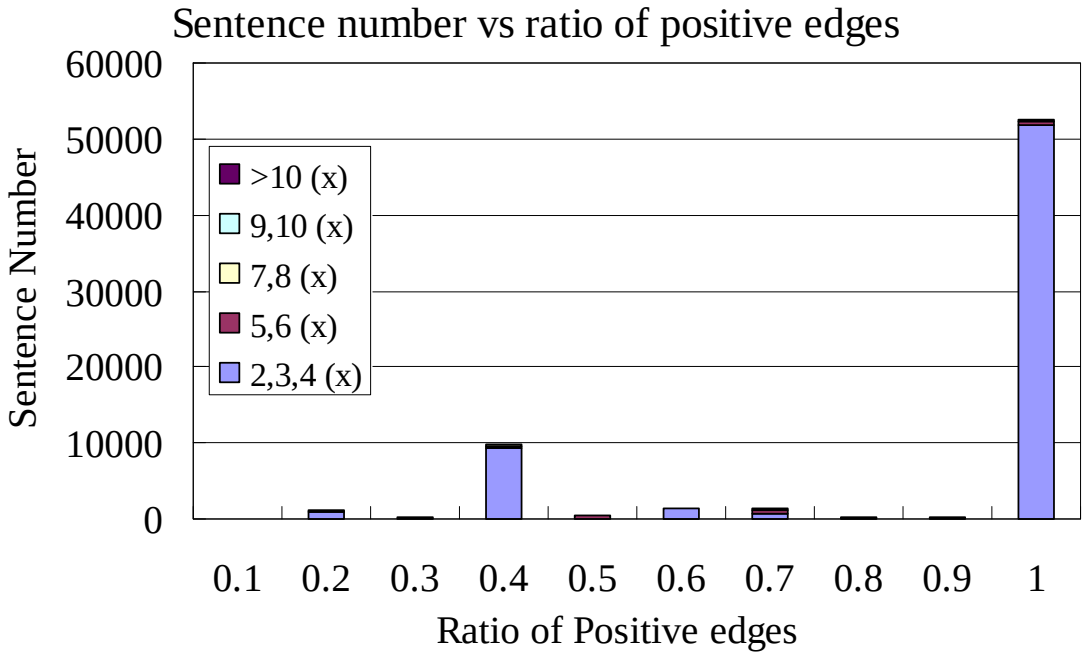


Part 18. Segmented: Molecular Function

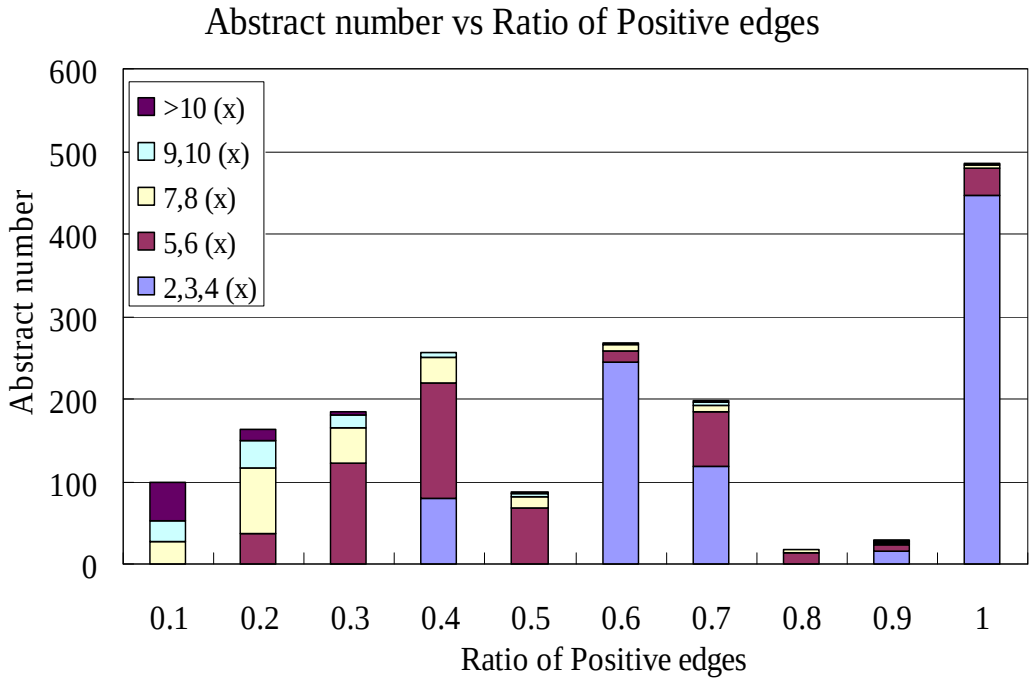
APPENDIX C



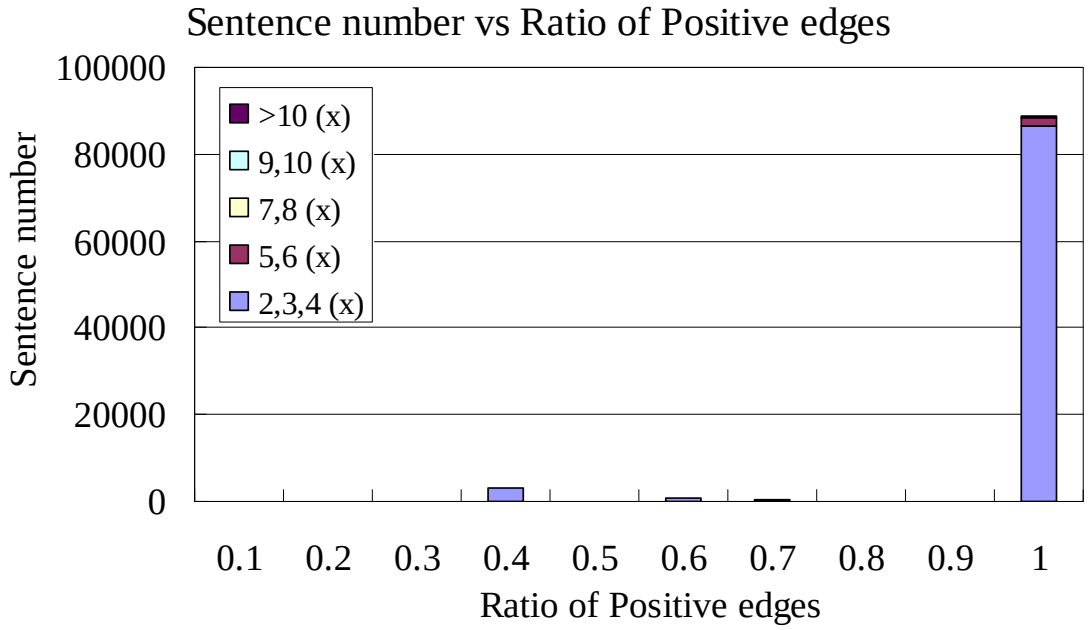
Part 1. Abstract number versus ratio of positive edges: **Cellular Component**



Part 2. Sentence number versus ratio of positive edges: **Cellular Component**



Part 3. Abstract number versus ratio of positive edges: **Molecular Function**



Part 4. Sentence number versus ratio of positive edges: **Molecular Function**

APPENDIX D

Ratio of Positive edges	No of Abstract with (x) Protein in abstract				
	2,3,4 (x)	5,6 (x)	7,8 (x)	9,10 (x)	>10 (x)
0.1	0	0	76	60	74
0.2	0	95	144	52	49
0.3	0	294	107	28	25
0.4	231	216	61	16	12
0.5	0	215	52	12	7
0.6	446	62	31	18	7
0.7	354	223	11	10	6
0.8	0	43	31	5	4
0.9	54	28	8	0	3
1	1223	138	37	11	1

Part 1. Abstract analysis of number of abstract with respective ratio of positive edges
GO domain: Biological Process

Ratio of Positive edges	No of Sentence with (x) Protein in sentence				
	2,3,4 (x)	5,6 (x)	7,8 (x)	9,10 (x)	>10 (x)
0.1	0	1	1	0	0
0.2	70	7	2	2	0
0.3	0	13	8	0	0
0.4	2540	26	0	3	0
0.5	0	82	7	4	0
0.6	487	0	16	0	0
0.7	84	219	1	7	1
0.8	0	12	47	9	4
0.9	15	16	3	6	12
1	90299	1912	295	79	38

Part 2. Sentence analysis of number of sentence with respective ratio of positive edges
GO domain: Biological Process

Ratio of Positive edges	No of Abstract with (x) Protein in Abstract				
	2,3,4 (x)	5,6 (x)	7,8 (x)	9,10 (x)	>10 (x)
0. 1	0	0	76	66	76
0. 2	0	65	145	46	25
0. 3	0	250	79	21	12
0. 4	92	178	24	13	4
0. 5	0	96	13	3	3
0. 6	391	16	10	2	1
0. 7	287	118	9	3	0
0. 8	0	12	12	1	0
0. 9	22	8	1	3	1
1	887	60	9	1	2

Part 3. Abstract analysis of number of abstract with respective ratio of positive edges
GO domain: Cellular Component

Ratio of positive edges	No of Sentence with (x) Protein in sentence				
	2,3,4 (x)	5,6 (x)	7,8 (x)	9,10 (x)	>10 (x)
0.1	0	13	13	4	0
0.2	935	115	33	6	4
0.3	0	257	46	10	9
0.4	9442	202	48	17	6
0.5	0	419	53	16	6
0.6	1305	25	36	7	6
0.7	761	517	10	7	7
0.8	0	57	59	8	5
0.9	92	80	18	13	6
1	51811	578	63	22	6

Part 4. Sentence analysis of number of sentence with respective ratio of positive edges
GO domain: Cellular Component

Ratio of Positive edges	No of Abstract with (x) Protein in Abstract				
	2,3,4 (x)	5,6 (x)	7,8 (x)	9,10 (x)	>10 (x)
0. 1	0	0	27	26	47
0. 2	0	36	81	32	15
0. 3	0	123	42	16	3
0. 4	79	141	31	5	0
0. 5	0	68	14	4	1
0. 6	245	14	7	1	1
0. 7	119	65	9	3	2
0. 8	0	13	5	0	0
0. 9	16	8	1	3	1
1	446	33	4	0	3

Part 5. Abstract analysis of number of abstract with respective ratio of positive edges
GO domain: Molecular Function

Ratio of Positive edges	No of Sentence with (x) Protein in sentence				
	2,3,4 (x)	5,6 (x)	7,8 (x)	9,10 (x)	>10 (x)
0. 1	0	5	5	0	0
0. 2	102	15	2	2	0
0. 3	0	14	7	0	0
0. 4	2910	30	2	1	0
0. 5	0	91	8	4	0
0. 6	581	1	18	0	0
0. 7	32	276	1	6	3
0. 8	0	4	52	17	1
0. 9	1	0	2	9	13
1	86634	1838	283	71	38

Part 6. Sentence analysis of number of sentence with respective ratio of positive edges
GO domain: Molecular Function