BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(16)

A text mining approach to detect mentions of protein glycosylation in biomedical text

Daksha Shukla[1] & Valadi K Jayaraman[2]*

[1]Bioinformatics Centre, University of Pune, India; [2]Centre for Development of Advanced Computing, University of Pune, India; Valadi K Jayaraman – Email: jayaramanv@cdac.in; Phone: 91-20-25704228; *Corresponding author

Received July 18, 2012; Accepted August 03, 2012; Published August 24, 2012

Abstract:
Protein Glycosylation is an important post translational event that plays a pivotal role in protein folding and protein is trafficking. We describe a dictionary based and a rule based approach to mine 'mentions' of protein glycosylation in text. The dictionary based approach relies on a set of manually curated dictionaries specially constructed to address this task. Abstracts are then screened for the 'mentions' of words from these dictionaries which are further scored followed by classification on the basis of a threshold. The rule based approaches also relies on the words in the dictionary to arrive at the features which are used for classification. The performance of the system using both the approaches has been evaluated using a manually curated corpus of 3133 abstracts. The evaluation suggests that the performance of the Rule based approach supersedes that of the Dictionary based approach.

Key words: Text mining, Glycosylation, Rule-based approach, Dictionary -based approach

Background:
Protein glycosylation is the most common post[7] translational modification of proteins. It is a complex process involving many functional proteins resulting in a great diversity of carbohydrate–protein bonds and glycan structures. Glycosylation plays a key role in biological processes and is linked to several molecular and genetic disorders. Glycosylation of some proteins has a great impact on their structures and functions resulting in modulation of many important biological processes. Alterations in glycosylation occur in many pathological states, and genetically determined defects in glycosylation are the reason of severe diseases. Thus, literature on protein glycosylation has been of interest to several researchers. However, the ever increasing amount of scientific literature and biological data calls for tools and methods to make this information available from text into more computable forms. Existing approaches to annotate the data in biological databases rely heavily on expert human curation **[1].** Given the growing volume of literature and new high-throughput methods, it is becoming necessary to provide tools that can reduce time and cost of curation, increase consistency of annotation, and provide the linkages to supporting evidence in the literature that make the annotations useful to researchers. Several dictionary and rule based approaches have been designed in the past decade to address similar needs. Dictionary based approaches have been extensively used in text mining systems. Oscar is an open source system for recognizing chemical entity 'mentions'; it integrates a dictionary of compound names, as well as using regular expressions, heuristics, and certain word combinations to find chemical names in text. Several recent works include dictionary look up: PLAN2L, a web-based online search system that integrates text mining and information extraction techniques **[2].** Enzyme databases such as FRENDA and AMENDA are additional databases created by continuously improved text-mining procedures and employ the usage of dictionaries **[3].** MGI scans full text articles and performs entity recognition for mouse gene 'mentions' based on a dictionary of mouse genes and human orthologs **[4].** Also, several teams at the Biocreative task **[5]** have also employed dictionary based approaches for term identification

Rule based approaches are frequently used for text mining tasks. One of the most successful rules-based approaches to gene and protein NER (Named Entity Recognition) in biomedical texts has been the AbGene system of Tanabe and Wilbur. ProMiner **[6]** is a dictionary- and rule-based system that applies sophisticated algorithms for recognizing complex, multi-word named entities in abstracts and full text articles. Other methods include statistical and machine learning based methods such as Support Vector Machines, co-occurrence based approaches and C-value method. With this work we describe a text mining approach to categorize text that has 'mentions' of protein glycosylation from those that do not. The work presented here employs both a dictionary based as well as a rule based approach to detect abstracts having 'mentions' of glycosylation. The dictionary based approach was designed keeping in mind that certain concepts can be expressed using different type of grammar. However, the choice of usage of terms related to glycosylation is restricted to a limited set of words and their synonyms. The dictionary based approach tries to cover the scope of these words that one can use if possibly speaking of or referring to protein glycosylation. The abstracts were screened for 'mentions' of these terms which have been categorized into 10 different dictionaries and then scored using several scoring schemes. For each scoring scheme a weighting methodology was used so that a different confidence is associated with each type of word as all the terms in the dictionary are not equally important with respect to glycosylation. Also these terms might be associated with text that is not related to glycosylation or merely as a passing reference. For example an abstract may speak of homology modeling of mucin (a glycosylated protein which is a part of the dictionaries constructed). However the text may be irrelevant with respect to glycosylation. This calls for the development of a robust scoring and thresholding scheme to distinguish such text from the relevant one. The rule based approach makes use of the J48 module available at the WEKA data mining tool **[7].** J48 is an implementation of the C4.5 algorithm for generating a pruned or unpruned decision trees. A set of words having high information content were selected for framing these rules. Several combinations were tried to arrive at a set of rules with the best coverage and support. The performances of both the approaches, as well as variations within the approaches have been evaluated using a manually curated corpus.

Methodology:

A summary of the methodology incorporated has been diagrammatically represented **(Figure 1).**

Preparation of the Corpora

Abstracts were fetched from PubMed by querying for the terms associated with protein glycosylation and subjected to the following process: i) Abstracts were annotated as positive if they contained information on the glycosylation process along with at least a 'mention' of the glycoprotein, glycosylating enzyme or the site of glycosylation; ii) Abstracts having some 'mentions' from the dictionaries however not implying glycosylation were selected as the negative dataset. Two datasets were prepared in this fashion. Corpus 1 which comprises of 300 positive & 300 negative abstracts was used in the initial studies to check for the ability of the approach to categorize the dataset. In order to validate the results obtained the dataset was scaled up to contain 1600 positive and 1533 negative abstract Dictionary based approach

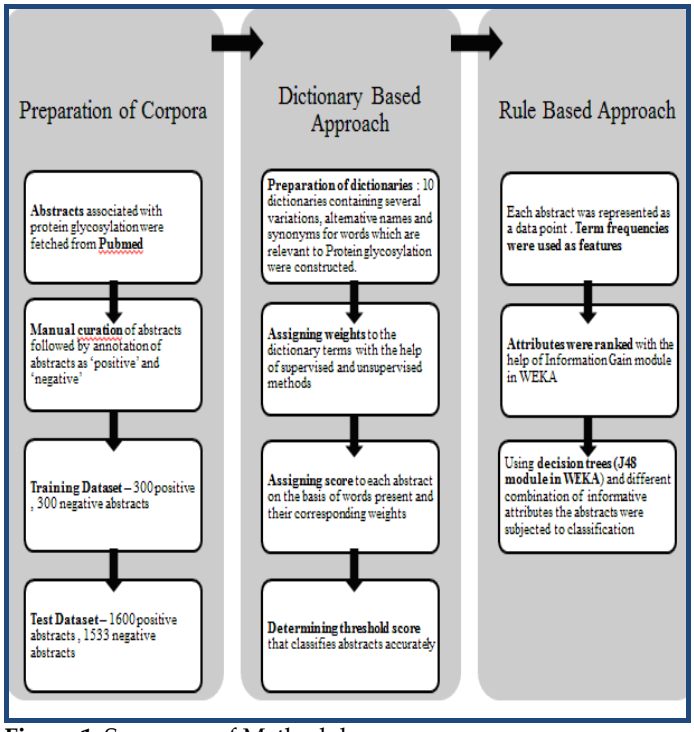

Figure 1: Summary of Methodology

Preparation of Dictionaries

10 dictionaries containing several variations, alternative names and synonyms for words which are relevant to Protein glycosylation were constructed: i) Carbohydrate names; ii) Amino acid common names, one letter code and three letter codes and chemical formula; iii) Gene Ontology Terms; iv) Pathway Names; v) Names and synonyms of enzymes that glycosylate proteins from several resources such as SwissProt, TrEMBL, BRENDA, CAZY and KEGG BRITE databases; vi) Gene Names and synonyms of enzymes that glycosylate proteins obtained from the alternative names and recommended gene names section of the SwissProt database; vii) E.C numbers of enzymes those glycosylate proteins; viii) Names and synonyms of proteins that are observed/reported to have undergone glycosylation from the UniProt database, O-GLYCBASE and MESH database; ix) Gene Names and synonyms of proteins that are observed/reported to have undergone glycosylation from SwissProt entries; x) Words that are observed to appear most frequently in biomedical text having 'mentions' of protein glycosylation with the help of LingPipe tool.

Derivation of weights

Weighting schemes were devised in order to attach higher significance to dictionary terms that are strong markers of glycosylation vis-a-vis ones that are not. The Weighting methodologies used in this approach are those that are proposed by Lan *et al* **[8]** and are classified into two types.

Unsupervised weighting methods

This approach included using normalized weights (weight for the dictionary depends on the contribution of words from that particular dictionary with respect to the total set of words from

all the dictionaries), term frequency (TF), normalized term frequency (TF values are normalized with respect to the number of words in the abstract), TF-IDF as used by Lan *et al* **[8]** as the weights.

Supervised weighting methods

This approach included using Relevance Factor (RF), TF-RF, and Normalized TF–RF as the weights. The weighting schemes in the Supervised category are also implemented as that used by Lan *et al* **[8].**

Scoring of abstracts and determination of thresholds

PERL scripts were written to match occurrences of terms from the dictionary. The frequency of occurrence of terms in combination with the weights associated with dictionary to which they belong was used to compute a score for the abstracts. A combination of thresholds was used to arrive at a threshold that could accurately categorize abstracts. Once the abstracts were scored using all the schemes the results were then analyzed to assess the performance by means of statistical measures like Precision, Recall, and F-measure.

Rule based approach

J48 module available at the WEKA data mining tool was used for rule based approach.

Feature Calculation

In the rule based approach used, individual abstracts were considered as data points. The features for these abstracts were given as numerical values which were vectors of fixed lengths. Features used were the term frequencies in the abstracts.

Feature Selection

A subset of high frequency words in the corpora were used to select the most informative attributes. On subjecting them to the Information Gain module for the ranking of attributes, the J48 algorithm returns a tree which contains decision rules. Rules were selected so as to contain only those with good Coverage and Support. Instances which were not covered by any rule were assigned to majority class of the left out examples.

Approaches using Decision Rules

Approach 1: Subset of 100 features selected by Information Gain; **Approach 2**: Since most of the terms in these 100 features come from the 'most-frequently occurring words dictionary' an approach which filtered out these words was used; **Approach 3:** Subset of 500 features from the dictionary words excluding words from the most frequent dictionary; **Approach 4**: The term occurrences were normalized with respect to the occurrence of terms from the dictionary to which the term in question belong to.

Discussion:

It was observed that the weighting methodologies were not able to assign very different weights to the different dictionaries. This resulted in similar resolving power for terms from all the dictionaries and a relatively poor performance of the dictionaries as compared to the rule based approach and the machine learning approach. Since the normalized weights did have some significant differences for different dictionaries, this approach fared better than other weighting methodologies. Several rule based approaches were tested for performance. In the first approach the attributes used as input comprised of frequencies of the top 100 high frequency words from the dictionaries in individual abstracts. Since the features in the above approach were independent of the dictionary from which the terms came from, another approach was devised such that dictionary information was also taken into account. Thus, in the second experiment the attributes used were the frequencies normalized with respect to the frequency of occurrence of the words from the dictionary from which the word came. It was observed that on normalizing the frequencies of occurrence of the words from the dictionaries with respect to the dictionary from which the word came there was no significant increase in the accuracy of categorization **Table 1 (see supplementary material).** Since the approaches used so far resulted in maximum number of rules coming out of the "Most Frequently" occurring words dictionary another approach was devised. In this approach the same procedure was repeated as before. But this time the terms from the most frequent words Dictionary were not included in the feature set. This was done in order to arrive at some domain related rules from other dictionaries whose performance was being over shadowed by the "Most Frequently" occurring words dictionary. However, the removal of these words from the attributes resulted in a significant drop in accuracy for both the corpora, the results for which are tabulated. All the approaches discussed above were repeated, such that numbers of features selected were scaled from 100 to 519. This approach performed well when 'Most frequent words' dictionary was included. However on removal of the attributes from the 'Most frequent words' dictionary this approach too showed poor performance in spite of significantly increasing the words from the other dictionaries. Thus the strength of the dictionary approach lies with terms from the most frequently occurring word dictionary.

Conclusion:

With the enormous increase in Scientific literature, researchers and manual curators are unable to keep at pace with the available information. There is thus a pressing need for text mining systems that automate the process and simplify this task. Several text mining systems have been successfully developed and tailored to the several problems at hand. With this study we have tried to use a text mining approach in order to detect mentions of protein glycosylation in text. The study has been conducted only with biomedical abstracts. The approach has used both a Dictionary-based and a Rule-based approach for this task. Categorical dictionaries have been created for this task. These dictionaries serve as the reference for the term identification in text. In the Dictionary-based approach a number of schemes for scoring have been experimented with. Both the supervised and unsupervised weighting do not show any significant difference in performance. The "Rule-based approach" fares better than the "Dictionary-based" for both the Corpora used for this task. Reduction in the dictionary size shows some improvement in performance.

Acknowledgement:

Dr. VKJ gratefully acknowledges DST, New Delhi for financial support and we also thank CDAC, Pune and Department of Bioinformatics, University of Pune for providing the required facilities to carry out our research work.

References:
[1] Altman RB *et al*. *Genome Biol*. 2008 **9**: S7 [PMID: 18834498]
[2] Krallinger M *et al*. *Nucleic Acids Res*. 2009 **37**: W160 [PMID: 19520768]
[3] Chang A *et al*. *Nucleic Acids Res*. 2009 **37**: D588 [PMID: 18984617]
[4] Dowell KG *et al*. *Database (Oxford)*. 2009 **2009**: bap019 [PMID: 20157492]
[5] Cohen AM & Hersh WR, *Brief Bioinform*. 2005 **6**: 57 [PMID: 15826357]
[6] Blaschke C *et al*. *BMC Bioinformatics*. 2005 **6**: S16 [PMID: 15960828]
[7] Tanabe L & Wilbur WJ, *Bioinformatics*. 2002 **18**: 1124 [PMID: 12176836]
[8] Lan M *et al*. *IEEE Trans Pattern Anal Mach Intell*. 2009 **4**: 721 [PMID: 19229086]

Edited by P Kangueane
Citation: Shukla & Jayaraman, Bioinformation 8(16): 758-762 (2012)

BIOINFORMATION

open access

Supplementary material:

Table 1: Summary of Results

DICTIONARY BASED APPROACH								
Metric	**Corpus**	**NW**	**TF**	**N-TF**	**RF**	**TF-RF**	**N-TF-RF**	**TF-IDF**
Accuracy (%)	1	81	74.8	72.16	75	75	72.5	72
	2	72.2	68.04	68.3	70.31	70.34	69.13	67.6
Precision (%)	1	79.6	73.8	71.8	74.03	74.6	72.57	71.15
	2	70.77	69.3	69.2	70.93	71.63	70.48	67.2
Recall (%)	1	83.33	77	72.33	77	75.66	72.33	74
	2	77.81	67	68.56	70.93	69.43	68.06	71.6
F-measure (%)	1	81.42	75.36	72.06	75.48	75.12	72.37	72.54
	2	73.9	68.13	68.87	70.93	70.51	69.24	69.33

NW: Normalized Weights ; **TF:** Term Frequency; **N-TF:** Normalized Term Frequency; **RF:**Relevance Factor; **TF-RF**: Term Frequency-Relevance factor; **N-TF-RF**: Normalized Term Frequency-Relevance factor; **TF-IDF**: Term Frequency-Inverse Document Frequency

RULE BASED APPROACH (Training Data)			
Approach	**Attributes**	**Correctly classified**	**Incorrectly classified**
100 FEATURES	Normalized frequencies	92.68%	7.32%
	Frequencies alone	93.18%	6.82%
100 FEATURES (EXCLUDING TERMS FROM MOST FREQUENTLY OCCURRING WORDS DICTIONARY)	Normalized frequencies	66.22%	33.77%
	Frequencies alone	64.89%	35.10%
519 FEATURES	Normalized frequencies	93.70%	6.30%
	Frequencies alone	94.34%	5.66%
519 FEATURES (EXCLUDING TERMS FROM MOST FREQUENTLY OCCURRING WORDS DICTIONARY)	Normalized frequencies	65.33%	34.67%
	Frequencies alone	66.06%	33.94%

RULE BASED APPROACH (Test Data)			
Approach	**Attributes**	**Correctly classified**	**Incorrectly classified**
100 FEATURES	Normalized frequencies	94.16%	5.84%
	Frequencies alone	94.03%	5.97%
100 FEATURES (EXCLUDING TERMS FROM MOST FREQUENTLY OCCURRING WORDS DICTIONARY)	Normalized frequencies	64.12%	35.87%
	Frequencies alone	64.76%	35.23%
519 FEATURES	Normalized frequencies	96.90%	3.09%
	Frequencies alone	96.95%	3.05%
519 FEATURES (EXCLUDING TERMS FROM MOST FREQUENTLY OCCURRING WORDS DICTIONARY)	Normalized frequencies	65.94%	34.05%
	Frequencies alone	66.06%	33.94%