An Analysis of Rule based Mining for Protein Phosphorylation

S.Padmapriya[1] R.Indra[2] J.Sengottuvelu[3]
[2]Assistant Professor
[1,2]Department of Computer Applications
[1,2]Shrimati Indira Gandhi College, Tiruchirapalli [3]Value Tech

Abstract— The study of proteins and its structure is a vast and complex subject. Earlier there has been lots of effort to classify and categorize the proteins structure. The focus of the current study is to generate rules based on information extraction models using data mining techniques. The architecture should not have computational overheads and the rule-based Information Extraction engine should implement all the features and display the patterns in a consistent mode. In normal information extraction networks, it tends to transmit rules in response to extracellular protein structure stimuli and other intracellular balance changes. The current work focuses on protein phosphorylation information, but the IE pipeline based model architecture can be instantly ported to the extraction of types other than phosphorylation. The rules generated are shown as graphs for analysis purposes.

Key words: Rule Mining, Protein Phosphorylation, Association Rules

I. INTRODUCTION

A. *Protein Phosphorylation*

Protein phosphorylation is a post-translational modification of proteins. The process consists where an amino acid residue is phosphorylated using a protein kinase. This is done by adding a covalently bound phosphate group. The process of phosphorylation alters the protein structure. This causes activated the functions by deactivating or by modifying. The opposite reaction of phosphorylation is called dephosphorylation. It is catalyzed by protein phosphatases. Both protein kinases and phosphatases work independently in order to balance or regulate the protein functions. First Protein phosphorylation was reported by Phoebus Levene in 1906 at the Rockefeller Institute for Medical Research and it took 50 years to discover protein kinases.

The problem in focus here is that a rule-based Information Extraction system, which extracts separately the kinase, the substrate, and the site. The information extraction model utilizes the lexical, syntactic, and semantic patterns of Protein structure. The problem to be solved is how new patterns are extracted. All the identified patterns would be present in the structure. However one main difference of phosphorylation from structure is to use a different name by which the new entity is recognized. The proposed model architecture applies a general filter and a substrate mode to identify and extract relevant features which are obviously sentences and site mentions when extracting relevant rules.

II. RELATED WORK

Z. Z. Hu1 et al proposed the RLIMS-P and the model showed precision recall at 91.4 and 96.4% respectively for paper retrieval, and the following accuracy of 97.9 and 88.0% for the extraction of substrates and sites. RLIMS-P had a high recall for paper retrieval with precision for information extraction for documents in protein phosphorylation. Jin-Dong Kim et al in their work proposed BioNLP'09 Shared Task along with reports, results of the analysis, including shared tasks with three sub-tasks, each of which addressed the bio molecular event extraction at specific levels. Yun Xu et al proposed MinePhos which is a SVM-based system and it outperformed RLIMS-P in both precision and recall of document information extraction on articles from PubMed. Manabu Torii et al proposed a modified PubMed Central (PMC) Open Access Subset, and obtained favourable results in mining full-text articles. Russ B Altman et al investigated several broad themes which included the possibility of fusing literature and biological databases through text mining.

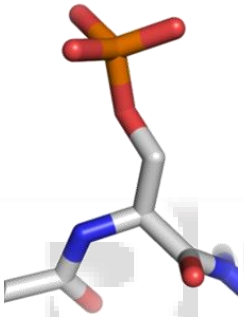

Fig. 1: Protein Phosphorylation

They tailored user defined classes of users and supported the scaling text mining technology by inserting it into larger workflows. L Hirschman et al reviewed recent results in literature data mining specifically for biology and proposed the steps for a challenge in evaluation for this field by creating challenging evaluations. Raul Rodriguez-Esteban et al proposed models where text mining applications are tailored to specific types of text and for considered reasons of cost and availability. M. S. Simpson et al provided a brief overview of the text mining done in biomedical domain and proposed the major text mining tasks like recognition of explicit facts from biomedical literature. They discovered previously unknown or implicit facts like document summarization and included question and answering texts.

A. *Protein Dataset*

The protein structure dataset is downloaded from the UCI repository. The dataset is cleaned and loaded into the model in this module. This contains attributes to be classified and clustered. Chromosomal coordinates of protein coding sequences - CDS only, or exons Predicted molecular weight (kDa), predicted pI, Charge and lengths (residues) on the protein data ignore the first ten or so numbers, they are parameters for the networks. The dataset uses '<' which is a spacer between each of the proteins and marks both the beginning and the end of the structure. The three characters namely - GLY are the 3 character codes used to identify the 20 amino acids. In this the character EH - E stands for beta-

sheet, the H for helix and the _ for 'other' or 'coil'. Always the biophysical constants number are ignored.

B. Cross Validation

Cross-validation is the first rule mining model validation technique used here for primarily assessing the results of the extracted protein structure from independent data set. It generates the goal prediction model for protein phosphorylation. The training dataset of known protein data and the test data are fed into the cross validation algorithm of weka. The desired goal of the cross validation is obtained in the training phase and the overfits are identified.

C. Visualizer Graph

The minimum spanning graph cluster is based on the applied rule based clustering algorithm. Here there are nil assumptions about the data points because the points are grouped at the centers and separated by a regular geometric curve. This step involves partitioning the datasets and select representative features. In correlation the relevant correlation measures are applied, hence the irrelevant features are removed. After removal the graph is then composed of the two connected components which in turn rid the dataset of irrelevant features and the process is called redundant feature elimination. The process is to extract only the features relevant to the target concept and to eliminate the irrelevant ones. The latter removes redundant features from relevant ones by choosing representatives from different feature clusters. This produces the final subset.

III. PERFORMANCE EVALUATION

The datasets for testing and benchmarking the mining system were derived from data sources in UCI Repository. Specifically, the annotation tagged protein models were used. These were developed for evidence attribution and phosphorylation features were annotated in the files. .

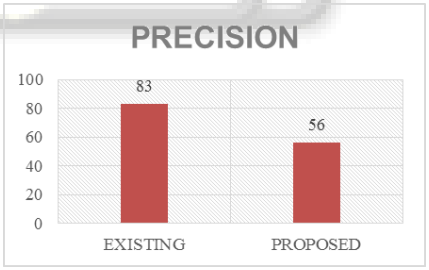

Fig. 2: Precision

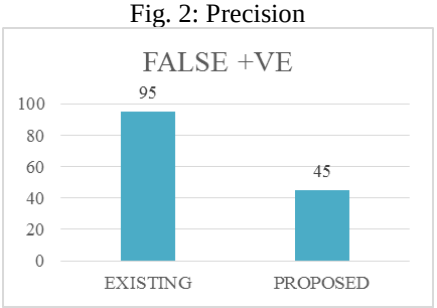

Fig. 3: False +ve

The system was evaluated for Information Rule extraction performance in two stages. One that the training dataset protein model was applied and then the 30% was used as the test dataset. Even that model was good and the factors used retrieved lots of rules for classification accuracy. The precision factor was overall good.

Four positive ones were missed, yielding a recall of 96.4%. The major improvement of system performance over the earlier models was mainly due to the addition of new diverse patterns, especially those containing phospho-residues.

Fig. 4: True +ve

IV. CONCLUSION

The designed model is an enhanced, generalizable architecture of a rule-based IE engine, and it works well for protein phosphorylation information extraction. The model architecture can be used for training and evaluating phosphorylation IE systems. The analysis solved the problem by covering a diverse Information Extraction pattern of protein phosphorylation. Additionally, a large set of text collection from full-text articles was annotated. The proposed model showed uniformly good performance across multiple criteria, and the protein phosphorylation rule extraction models. This shows that the model is robust and scales well to phosphorylation Information extraction from diverse protein data sets containing full attribute specifications. In future the partition cluster rules can be made interactive by supplying values to the nodes in an interactive and dynamic manner. Dynamic configuration sets from live real data will be used by such a model and will enhance user perception in the future.

REFERENCES

[1] T. Hunter, "Why nature chose phosphate to modify proteins," Philos. Trans. Royal Soc. London B. Biol. Sci., vol. 367, no. 1602, pp. 2513–2516, Sep. 2012.

[2] D. A. Natale, C. N. Arighi, J. A. Blake, C. J. Bult, K. R. Christie, J. Cowart, P. D'Eustachio, A. D. Diehl, H. J. Drabkin, O. Helfer, H. Huang, A. M. Masci, J. Ren, N. V. Roberts, K. Ross, A. Ruttenberg, V. Shamovsky, B. Smith, M. S. Yerramalla, J. Zhang, A. Aljanahi, I. Celen, C. Gan, M. Lv, E. Schuster-Lezell, and C. H. Wu, "Protein Ontology: A controlled structured network of protein entities," Nucleic Acids Res. , vol. 42, pp. 415–421, Nov. 2013.

[3] P. V. Hornbeck, J. M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, V. Latham, and M. Sullivan, "PhosphoSitePlus: A comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse," Nucleic Acids Res., vol. 40, no. Database issue, pp. D261–270, Jan. 2012.

[4] H. Dinkel, C. Chica, A. Via, C. M. Gould, L. J. Jensen, T. J. Gibson, and F. Diella, "Phospho.ELM: A database of phosphorylation sites–update 2011," Nucleic Acids

Res., vol. 39, no. Database issue, pp. D261–D267, Jan. 2011.

[5] The UniProt Consortium, "Reorganizing the protein space at the Universal Protein Resource (UniProt)," Nucleic Acids Res., vol. 40, no. Database issue, pp. D71–D75, Jan. 2012.

[6] Z. Z. Hu, M. Narayanaswamy, K. E. Ravikumar, K. Vijay-Shanker, and C. H. Wu, "Literature mining and database annotation of protein phosphorylation using a rule-based system," Bioinformatics vol. 21, no. 11, pp. 2759–2765, Jun. 2005.

[7] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, "Overview of BioNLP'09 shared task on event extraction," in Proc. Workshop BioNLP: Shared Task, 2009, pp. 1–9.

[8] M. Narayanaswamy, K. E. Ravikumar, and K. Vijay-Shanker, "Beyond the clause: Extraction of phosphorylation information from medline abstracts," Bioinformatics, vol. 21, o. Suppl 1, pp. i319–327, Jun. 2005.

[9] A.-L. Veuthey, A. Bridge, J. Gobeill, P. Ruch, J. R. McEntyre, L. Bougueleret, and I. Xenarios, "Application of text-mining for updating protein post-translational modification annotation in UniProtKB," BMC Bioinformatics, vol. 14, no. 1, p. 104, Mar. 2013.

[10] Y. Xu, D. Teng, and Y. Lei, "MinePhos: A literature mining system for protein phosphorylation information extraction," IEEEACM Trans. Comput. Biol. Bioinformat., vol. 9, no. 1, pp. 311–315, Apr. 2011.

